Construction of Pronunciation Instructions on Tongue and Lips Movements based on Formants of Speech Signal

Kyoko Yoshiwara*, and Izumi Hanazaki*

* Department of Robotics and Mechatronic, School of Science and Technology for Future Life,
Tokyo Denki University, 5, Asahi-cho Senju Adachi-ku Tokyo, 120-8551,Japan
TEL:+82-3-5284-5609,FAX:+81-3-5284-5698, e-mail: hana@fr.dendai.ac.jp

Abstract

Currently, we are developing English pronunciation training system for Japanese adults. In this system, pronunciation is indicated by displaying formants, lip shapes, and tongue shapes that are effective for identifying vowels. In this work, focusing on the visualization of tongue shape and lip shape, analysis of MR image and speech signal was performed. Generally, there is a correspondence between the first formant and the opening length of the mouth. Therefore, we obtained correlation coefficients of the first formant detected from speech signal and opening of lips and jaws detected from MR image. As a result, it was found that some cases showed strong correlation or not depending on subjects. From past research, it is known that the Japanese vowel /u/ which is the lingual tongue vowel may change to the front tongue due to the protrusion of the lip. From this, it is considered that the subject who don’t show strong correlation manipulates the phoneme by projecting the lips instead of opening the mouth vertically. In order to support this hypothesis, it is necessary to analyze more MR images and speech signals in the future.

1. Introduction

One of the problems when adults acquire non-native languages is that articulatory organs such as tongue and lips are optimized for mother tongue pronunciations, so it is difficult to make non-native pronunciation actions. Therefore, in case Japanese people acquire English, accurate instruction of tongue and lip motion is necessary. According to the questionnaire about Japanese education of English, more than half of the elementary school teachers answered "I don’t have knowledge about English pronunciation" or "I don’t have confidence in my pronunciation". On the other hand, more than half of the high school teacher and universities lecturers answered that they are confident in their English pronunciation and knowledge. But practically, they don't make guidance of pronunciation much[1]. Therefore, we are developing English pronunciation training system which can instruct proper pronunciation motion on behalf of teacher. In this report, we analyze opening and closing of mouth and jaw when pronouncing Japanese vowel from speech signal and MR image, and clarify relationship between tongue movement and lip movement.

2. Current state of English pronunciation training system

2.1 Outline of pronunciation training system

In the currently developed system, teaching is performed by displaying the phoneme, lip shape, and tongue shape in real time by using the formant which is the feature quantity of the vowel phoneme as shown in Fig.1. Currently it corresponds only to English vowel, and correction is applied when the formant exceeds the range of English vowel.

Fig.1: English Pronunciation Training System

2.2 Movement of the highest point of the tongue surface by protrusion of the lip

In field of English, the tongue shape is an important element of articulatory control, so teaching about the position of highest point of tongue surface by the IPA chart is indispensable. However, from verification experiments using the system, it has proved difficult for the trainee to recognize its own tongue shape. Therefore we analyzed head MR image
and speech signal which was recorded at the same time, using reflection coefficient by vocal tract area function. As a result, it was found that the tongue surface highest point of the Japanese vowel / u / which is the posterior tongue vowel moves forward by protrusion of the lip[2]. From the above, it is necessary to consider habits of the trainee such as protrusion of lips.

2.3 Lip movement model based on facial muscle electromyography signal

Jannuzi has confirmed that teaching how to use the expression muscles of English speakers with facial muscle potential during pronunciation has an effect on pronunciation training[3]. Based on this fact, it is thought that the facial muscle electromyography signal is useful for our system. So we identified the relationship between the facial muscle of the English teacher who studied phonetics and the opening of the lip in the state space model, and created a lip motion model considering the cooperativeness of the multiple input system of the facial muscle. In this model, the masseter muscle, risorius muscle, orbicularis oris muscle, depressor anguli oris muscle, digastricus muscle shown in Fig.2 are input, and the longitudinal opening, lateral opening and protrusion of the lips shown in Fig.3 are output. The muscle potentiometer was measured with a myoelectric potentiometer, and the mouth opening was measured with motion capture.

Comparing the estimation of the model with the measurement result using the English word of Table.1 pronounced by the English teacher, words including vowels [Æ] and [e] are highly correlated, but words including [i] and [ʌ] are lower correlated. As an example Fig.4 shows estimation results and measurement results of the longitudinal opening, the lateral opening and protrusion of the mouth when the myoelectric potentials of the English words "set" and "sit" are input. In this way, the tendency of the correlation coefficient is divided by the type of vowel. So it can be seen that the relationship between the facial muscle electromyography signal and lip motion is influenced by the vowel rather than consonant [4].

Table 1. Correlation coefficient between model estimation result and measurement result

<table>
<thead>
<tr>
<th>Vowel/Word</th>
<th>Vertical Length</th>
<th>Horizontal Length</th>
<th>protruding</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ʌ]</td>
<td>0.042</td>
<td>0.066</td>
<td>-0.175</td>
</tr>
<tr>
<td>hut</td>
<td>-0.384</td>
<td>-0.188</td>
<td>0.020</td>
</tr>
<tr>
<td>[ɑ]</td>
<td>0.680</td>
<td>-0.175</td>
<td>-0.581</td>
</tr>
<tr>
<td>hot</td>
<td>0.257</td>
<td>0.458</td>
<td>0.125</td>
</tr>
<tr>
<td>cat</td>
<td>0.807</td>
<td>0.115</td>
<td>0.320</td>
</tr>
<tr>
<td>[æ]</td>
<td>0.439</td>
<td>0.050</td>
<td>0.413</td>
</tr>
<tr>
<td>sat</td>
<td>0.628</td>
<td>0.366</td>
<td>0.592</td>
</tr>
<tr>
<td>[i]</td>
<td>-0.125</td>
<td>-0.124</td>
<td>0.102</td>
</tr>
<tr>
<td>[e]</td>
<td>0.606</td>
<td>0.341</td>
<td>0.803</td>
</tr>
</tbody>
</table>

As mentioned in the previous section, some of the facial muscles affect vowels and others do not. In addition to

2.4 Visualization of tongue shape during pronunciation of English words

As mentioned in the previous section, some of the facial muscles affect vowels and others do not. In addition to
that, it seems that there are other factors that influence the pronunciation of vowels. We speculate that the movement of the tongue is a major factor in this problem, and are seeking to improve the lip model from the visualization model of the tongue shape which currently used in the system.

The tongue shape visualization model is designed to estimate the tongue shape from the formant of the speech signal uttered by the trainee. The formant is derived by an auto regressive (AR) model by the Burg method, and the tongue shape is estimated based on this formant and 5 MR images at the time of generation of a Japanese vowel by one subject.

As shown in Fig.5, since the vocal tract is hollow, it was displayed black in the MR image. Therefore, binarize MR image and extract the contour of the vocal tract to separate it into the palate and the tongue. Next, the center lines of the palate and tongue surface parts are determined. In finding the centerline, it is necessary to smooth the lips and dentition affecting it. Search for the point of the palate part located at the shortest distance with respect to any point on the tongue surface, and set the center of the two points it is obtained by making it a point on the center line. The center line is divided into 24, the intersection of the line orthogonal to the center line at the dividing point and the palate lid portion, the tongue surface portion is obtained, and the distance between the two points is taken as the characteristic amount of the tongue shape.

Fig.5: Extraction procedure of tongue position

We do multiple regression analysis to associate this feature amount with the formant. The multiple regression equation is shown in the following equation using the tongue shape parameter \( \hat{T}_j(j) \) as a dependent variable, the first formant \( F_1 \), and the second formant \( F_2 \) as independent variables.

\[
\hat{T}_j(j) = b + a_1F_1(j) + a_2F_2(j)
\]

Here, the tongue position feature quantities \( \hat{T}_j(j) \), \( b \) estimated from the formant are constant term and \( a_i[i=1,2] \) are partial regression coefficient.

The result of multiple regression analysis is shown in Fig.6. The adjusted coefficient of determination shows the goodness of fit of the multiple regression model, and the value is 0 for the vicinity of the middle of the tongue position feature quantity, but it is close to 1 otherwise. Actually, the vicinity of the middle does not move during pronunciation, so the coefficient of determination is to be 0. Therefore, the validity of the model is enough.

Fig.6: Result of multiple regression analysis

Using this model, we analyzed the shape of the tongue when the English teacher pronounced the words in Table 1, and found that there is similarity between the tongue shape of Japanese vowel and English vowel. In addition, many commonalities were found between the tongue shape visualized by the model and the conventional teaching method for pronunciation of the English vowels. From this, we think that teaching more intelligibly can be made by teaching gradually to approach the tongue shape of English vowels based on the tongue shape of Japanese vowels.

However, in this tongue shape visualization model, only the tongue shape is changed according to the pronunciation, and the degree of opening of the mouth is fixed irrespective of the pronunciation. Therefore, as shown in Fig.7, a tongue-shaped line sometimes penetrates the lower jaw. Accordingly, it is necessary to visualize the opening and closing of the mouth together with the tongue shape. For this reason, in this report, in order to improve the tongue shape visualization model, opening / closing of the jaw is detected from the MR image and the correspondence relation with the formant is clarified.
3. Detection of jaw position

3.1 Analysis conditions

We use the AR model which processed the frame under the condition of Table 2, and find the formant.

<table>
<thead>
<tr>
<th>Analysis conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sampling frequency</strong></td>
</tr>
<tr>
<td>AR order</td>
</tr>
<tr>
<td>Frame length</td>
</tr>
<tr>
<td>Shift length</td>
</tr>
<tr>
<td>subject</td>
</tr>
</tbody>
</table>

After that, the positions of the nose, upper lip, lower lip, and chin are manually detected from MR images which each subject was pronouncing 5 Japanese vowels and resting. Next, the position of the nose of waiting is aligned with the position of the nose of pronouncing the vowel, detects the Euclidean distance of the opening of the mouth or jaw from the upper lip to the lower lip (red line), from the upper lip to the chin (green line), from the nose to the lower lips (blue line), from nose to chin (yellow line).

In general, it is said that $F_1$ corresponds to the opening of the mouth, so we obtain the correlation coefficient between $F_1$ and mouth opening.

3.2 Results Identification

Fig.10 shows the MR image of each subject during rest. Since MRI was taken while laying down, there are differences such as different angle of face for each subject. In addition, there is also a possibility that big individual differences in the position and shape of the lip may occur at the time of rest and pronunciation due to factors such as tooth alignment.
The position of the upper lip, the lower lip, and the jaw from the subject 1 to the subject 4 are shown in Fig.11 to 14.

In subject 1, the movement width of each point in the horizontal axis direction is about 5 pixels, and the movement in the vertical axis direction is about 10 pixels. As a result, you can see that the overall motion is smaller than other subjects.

In subject 2, the movements of the jaws are small with respect to the movements of the upper lip, the width of the motion of each part of the horizontal axis is 20 pixels on the upper lip, and 10 pixels on the lower lip and jaw. Among them, it is obvious that protrusions of /u/ and /o/ are outstanding.

In subject 3, the upper lip didn’t move but the lower lip and jaw moved. The movement of the lower lip on the vertical and horizontal axis is about 10 pixels, but the jaw has a movement of 20 pixels in the vertical axis direction and about 10 pixels in the horizontal axis direction.

In subject 4, the upper lip of each vowel and the lower lip, the whole jaw is moving. The jaw moves as much as the movement of the upper lip. The movement width of each part on the horizontal axis is about 10 pixels of the upper lip, 15 pixels of the lower lip, and 13 pixels of the jaw. From the comparison with the resting MR image, it can be seen that the subject 4 is pronouncing by lowering the jaw while protruding the lips. The difference between the opening of the lower lip and the jaw at the time of standby and pronunciation can be said to be larger than the other subjects.

Next, Table 3 shows correlation coefficients between $F_1$ and mouth opening length of each subject.

<table>
<thead>
<tr>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>upper lip/ lower lip</td>
<td>0.284</td>
<td>0.500</td>
<td>0.604</td>
<td>0.878</td>
</tr>
<tr>
<td>upper lip/ jaw</td>
<td>0.225</td>
<td>0.558</td>
<td>0.724</td>
<td>0.819</td>
</tr>
<tr>
<td>nose/ lower lip</td>
<td>0.0669</td>
<td>0.481</td>
<td>0.723</td>
<td>0.906</td>
</tr>
<tr>
<td>nose/jaw</td>
<td>-0.0175</td>
<td>0.590</td>
<td>0.788</td>
<td>0.817</td>
</tr>
</tbody>
</table>

From Table 3, it can be seen that subject 1 has a weak correlation and subject 2 has a correlation, while subjects 3 and 4 have a strong correlation.

Subjects 1 and 2 know from the past study as shown chapter
2 section 2 that by originally projecting the lips, the articulation point of the Japanese vowel /u/ of the posterior tongue is used as the anterior tongue. Therefore, from the above, it can be considered that the subjects 1 and 2 are changing $F_1$ by protruding the lips instead of opening the mouth. From Fig.14, it appears that the subject 4 also protrudes the lips, but in the case of him, it seems that movement of the place of articulation did not occur because he protrude the lips while opening their mouth.

4. Conclusions

In this report, in order to examine improvement of the tongue shape visualization model used in the English pronunciation training system, the opening of the mouth and jaw was detected from the MR image and speech signal of the Japanese vowel. As a result, we found that there were some subjects who did not have strong correlation with subjects who had a strong correlation between mouth opening and $F_1$. In analysis of the past, subjects who don't have a strong correlation have been found to be person who have pronounced Japanese vowel /u/ with a front tongue, in spite of /u/ is generally pronounced with a back tongue. Therefore, it seems that these subjects are changing $F_1$ by thrusting out the lips instead of opening the mouth vertically.

Based on the above, it is necessary to analyze more MR images in the future, and to clarify the relationship between the shape of the tongue and the opening of the mouth in addition to the relationship between the opening of the mouth and the formant.

References


The MRI data and recorded sound data used in this study are parts of ”ATR vocal tract MRI data for Japanese vowels” that were acquired at Human Information Science Laboratories in Advanced Telecommunications Research Institute International (ATR) and released from ATR-Promotions Co. Ltd. The use of the data is under the license agreement with ATR-Promotions Co. Ltd.