Use of Reflection Coefficients of Burg’s Method for Improvement of Visual Support Way in Pronunciation Practice*

Ryosuke Karigome† and Izumi Hanazaki†

In pronunciation practice, the trainees essentially need interactive supports to progress. We have developed a pronunciation training system, where the lip posture and tongue shape that are estimated by the formants extracted by AR model using Burg’s method are shown to the trainee. Realization of real time operations is one of the most important problems in this system. However, formants estimation requires long computation time. In order to avoid spectrum estimation, we attempt to reconstruct these visual instructions from the reflection coefficients of Burg’s method. Here, we investigate relationship between the acoustic feature or articulation manner of the sound and the reflection coefficients. The vocal tract area function calculated from the reflection coefficient was confirmed to acquire feature of tongue shape extracted from MR image taken during uttering more precisely than the formant. This gave possibility to improve visual support way based on the formants.

1. Introduction

In ordinary circumstances, a speaker is not conscious of the processes involved in sound production, because he/she has learned them naturally. When the speaker produces an unfamiliar sound, he/she controls his/her articulator as to correct the difference between the own utterance and the objective sound. This is called auditory feedback. However, if the speaker is not able to detect the difference, a teacher to correct the speaker’s pronunciation is required. Conventionally, a language teacher has been involved in it. In addition, teaching material available on computers has been developed. This is called Computer Assisted Language Learning (CALL). In pronunciation practice, the trainees essentially need interactive supports to progress. Realization of real time operations is one of the most important problems to be solved in CALL, in addition to development of effective support ways [1].

We have been developing the pronunciation training system based on speech signal analysis. In our system, visualization support is based on formants of speech signal. The formants are extracted from the estimated spectrum of the speech signal by AR model and this process requires long calculation time. In order to reduce calculation time, we attempt to reconstruct these visual instructions from reflection coefficients of Burg’s method. The reflection coefficients are simultaneously calculated in Burg’s algorithm and are thought to include some information about the vocal tract according to the similarity of structures of Burg’s method and the acoustic model of vocal tract. Here, we investigate relationship between the formants or tongue shape and the reflection coefficients using principle component analysis.

2. Overview of the Pronunciation Training System

Visual instructions of the pronunciation training system are shown in Fig. 1. As the visual assistance instruction of this system, the degree of mouth opening and the shape of the tongue are calculated from the formants which are estimated in every 60ms. The speaker can recognize the difference of phoneme from these and correct own pronunciation in real time.

3. Similarity between Lattice Structure of Burg’s Method and Kelly’s Speech Production Model

3.1 Reflection Coefficients of Burg’s Method

AR (autoregressive) model is represented by the following equation.

\[ x(n) + \sum_{k=1}^{p} a_k^{(p)} \cdot x(n-k) = e(n) \]  

(1)
Where \( x(n)|n=0,1,2,\cdots \) is stationary signal, \( p \) is AR model order, \( a_k \) is AR model coefficient, and \( e(n) \) is prediction error. AR model for the speech signal is estimated by Burg’s method. In Burg’s method, AR coefficients are estimated so as to minimize both of forward prediction error and backward prediction error which are represented in (2) and (3) respectively.

\[
fe^{(p)}(n) = x(n) + \sum_{k=1}^{p} a_k^{(p)} \cdot x(n-k) \quad (2)
\]

\[
be^{(p)}(n) = x(n-p) + \sum_{k=1}^{p} a_k^{(p)} \cdot x(n-p+k) \quad (3)
\]

AR coefficients are recurrently calculated using reflection coefficients. The reflection coefficients are expressed by the following equation.

\[
k_p = \frac{-\sum_{n=p}^{N-1} f_e^{(p-1)}(n) \cdot b_e^{(p-1)}(n-1)}{\sum_{n=p}^{N-1} \left[ b_e^{(p-1)}(n-1) \right]^2 + \left[ f_e^{(p-1)}(n) \right]^2} \quad (4)
\]

### 3.2 Kelly’s Speech Production Model

Vocal tract can be approximated as series connection of cylindrical tube with different cross-sectional area. It is called Kelly’s Speech Production Model. The reflection coefficient of Kelly’s Speech Production Model is caused by the area ratio of the neighbor cylindrical tube. Propagation of vocal tract sound wave can be represented by the reflection coefficients. Vocal tract area function is expressed by the following equation.

\[
A_n = \frac{1-k_n}{1+k_n} A_{n-1} \quad (n=1,2,\cdots,p) \quad A_0 = 1 \quad (5)
\]

where \( A_n \{n=1,2,\cdots,p\} \) are vocal tract area function coefficients, and \( k_n \{n=1,2,\cdots,p\} \) are the reflection coefficients. Starting point of (5), \( A_0 \), denotes lips opening. Vocal tract area from lips to glottis can be obtained by changing the ratio of the cross-sectional area adjacent to each other according to each reflection coefficient. In Kelly’s vocal tract model, PARCOR coefficient is equal to the reflection coefficient.[2] Since PARCOR coefficient is equal to the reflection coefficient in Burg’s method, the reflection coefficient of Kelly’s Speech Production Model is regarded to equal to the reflection coefficient of the Burg’s method from the viewpoint of mathematical model. It can be said that the reflection coefficients of Burg’s method might represent the characteristic of the vocal tract. We estimate the vocal tract area function by substituting the reflection coefficients of Burg’s method for equation (5).

### 4. Relationship between the Formants and the Reflection Coefficients

The relationship between formants of the speech signal and the reflection coefficients of Burg’s method is analyzed by multiple regression analysis. In this analysis, we adopt the 1st and the 2nd formants which represent phoneme as dependent variables, and the reflection coefficients as the independent variables. In our previous research, we have confirmed that the vocal tract area function for the oral cavity can distinguish the vowels[3]. The length of the oral cavity is about a half of the total length of the vocal tract from the lips to the glottis. As the order of the AR model is set 16, the numbers of the reflection coefficients and the divided sections of the vocal tract are 16. The independent variables are reduced to the first 8 reflection coefficients, because the first 8 coefficients are corresponding to mouth cavity and are deeply dependent on the phoneme.

Dataset of multiple regression analysis are shown in Table 1. The reflection coefficients and the formants are obtained from the speech signals which 10 Japanese men in their 20s utter three times for each vowel. Coefficient of determination of the result is shown in Table 2.

From Table 2, the reflection coefficients seem to be closely correlated to the 2nd formant, because the coefficient of determination of the 2nd formant is close to 1. However, the 1st formant is not enough corresponding to them.

In order to rebuild a visual support from the reflection coefficients in addition to the formants, we must make furthermore investigation about the relationship between the 1st formant and the reflection coefficients.

<table>
<thead>
<tr>
<th>Subjects</th>
<th>10 Japanese men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recording vowels</td>
<td>Japanese vowels</td>
</tr>
<tr>
<td></td>
<td>/a/ /i/ /u/ /e/ /o/</td>
</tr>
<tr>
<td>Samples</td>
<td>150 samples</td>
</tr>
<tr>
<td></td>
<td>(3 samples per each vowel)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 1 Data set of multiple regression analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
</tr>
<tr>
<td>Recording vowels</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Samples</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2 Results of multiple regression analysis of the reflection coefficients and formants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient of determination</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>0.653869</td>
</tr>
</tbody>
</table>
5. Investigation of Vowel Articulation using Principal Component Analysis of the Reflection Coefficients

We shall examine the relationship between the reflection coefficients and vowel articulation in order to introduce the reflection coefficients to our pronunciation training system. As in the former session, AR model with 16 order is used. Based on our previous research result that the vowels are determined by the reflection coefficients corresponded to the oral cavity and pharyngeal cavity where the vowels are characterized[3], we will analyze the first 8 reflection coefficients by principal component analysis (PCA). PCA is derived from equation (6) so as to minimize the loss amount. The data matrix used for the principal component are shown in equation (7). $x_{Nv}$ are the measured data, $N$ is the number of total frames of the speech signal, and $v$ is the variable of the 8 order of the reflection coefficients. The data matrix $x_{Nv}$ are standardized with the mean 0 and the variance 1. $b$ are the eigenvector of the correlation coefficient matrix of the data matrix $X_{Nv}$, and $Z$ is the principal component.

$$Z = X_{Nv} b$$  \hspace{1cm} (6)

$$X_{Nv} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1v} \\ x_{21} & x_{22} & \ldots & x_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{Nv} \end{pmatrix}$$  \hspace{1cm} (7)

$$b = \begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1v} \\ b_{21} & b_{22} & \ldots & b_{2v} \\ \vdots & \vdots & \ddots & \vdots \\ b_{N1} & b_{N2} & \ldots & b_{Nv} \end{pmatrix}$$  \hspace{1cm} (8)

Data sets of PCA are shown in Table 3. The speech signals of 5 Japanese men in 20s were divided to the 30ms frame and the reflection coefficients for each section were estimated by Burg's method. 2500 sets of the first 8 reflection coefficients are analyzed by PCA. The first and second principal components are extracted and shown in Fig. 2 as the scatter diagram of their scores while the cumulative proportion and the eigenvalues of the correlation matrix are simultaneously shown.

From scatter diagram of the principal component scores, we can see that the ranges of the 1st principal component score of the vowels seem to become larger in order that /o/, /a/, /u/, /e/, /i/. Japanese vowels, /o/, /a/, and /u/ belong to the back vowel and /e/, /i/ belong to the front vowel. It means that the highest point of the tongue of the vowel /o/, /a/, /u/ is in the back of the tongue and the highest point of /e/, /i/ is in the front. Considering these, we can say that the 1st principal component scores can represent the position of the highest point of the tongue in the oral cavity. The highest point of the tongue is predicted to move to the back in the oral cavity, as the 1st principal component score decreases. Generally, the 2nd formant represents the front or back position of the high point of tongue.

The tongue front or back position as well as the tongue height (high to low) are important to describe vowel articulation and it became obvious that the former may be described by the 1st principle components, but the latter is not yet explained. Then, we will analyze all 16 reflection coefficients using PCA. The results are shown in Fig. 3. As shown in Fig. 3, in the scatter diagram of the principal compo-

<table>
<thead>
<tr>
<th>Subjects</th>
<th>5 Japanese men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recording vowels</td>
<td>Japanese vowels /a/, /i/, /u/, /e/, /o/</td>
</tr>
<tr>
<td>Samples</td>
<td>2500 datas (100 datas per each vowel)</td>
</tr>
</tbody>
</table>
Karigome and Hanazaki: Use of Reflection Coefficients of Burg’s Method for Improvement of Visual Support Way in Pronunciation Practice

6. Visualization of the Tongue Shape by Vocal Tract Area Function

6.1 Estimation of the Tongue Shape using Vocal Tract Area Function

Story et al. analyzed the vocal tract area function extracted from the MR image using PCA and showed that the 1st principal component described effectively the position of the front and back of the tongue [7]. Based on this fact, we will attempt to estimate the tongue position using the vocal tract area function which is calculated using the reflection coefficients of Burg’s method by substituting the coefficients to equation (5).

Then, we analyze the vocal tract functions obtained for the dataset shown in Table 3 by PCA. According to investigate similar as the former section, we have confirmed that the 1st principal component represent mainly the front or back position of the tongue. Therefore we attempt to estimate the tongue shape that specifies the highest point, using the 1st principal component of the vocal tract area function. The vocal tract area function is restored by the 1st principal component, following next equation.

\[ x_n = x_n b_1 b_1^T \] (9)

The vocal tract area function of \( n \{1, 2, \cdots, N\} \) frames are \( x_n \), and the eigenvector of the first principal component are \( b_1 \).

The vocal tract is modeled as the sound tube which is consisted of several short sound tubes with the same length but with the different cross-sectional area. The vocal tract area function represents the cross-sectional area of the sound tubes from the lips to the glottis. Since the palate is not deformed during uttering, being small of the vocal tract area function of the tube means that the tongue surface is up and being large means that the surface is down. Therefore we will approximate the tongue shape by reversing the vertical-axis of vocal tract area function. The estimated shape of tongue during articulation of each Japanese vowel are shown in Fig. 4. In this figure, we can see the relative position of the highest point of the tongue in the oral cavity. In Fig. 4, horizontal axis denotes the distance on the line from the lip to the pharyngeal in the oral cavity. In addition, each line shows the estimated result by the voice of each frame of 5 subjects.

As shown in Fig. 4, tongue shapes estimated from vocal tract area function were divided into two types. One is the front tongue in which the highest point is near the lip, and the other is the back tongue. It is said that Japanese vowels, /i/ and /e/, are belong to front tongue, /a/, /o/, and /u/ are belong to back tongue. From Fig. 4, we can see that the estimations of the vowels except /u/ are similar to this fact but /u/ is estimated in a different tongue shape depended on the individual subject.
6.2 Evaluation and Discussion of the Estimated Tongue Shape

We will validate the estimated tongue shape from the reflection coefficients by comparing them to the tongue shape extracted from MR image which were taken during the vowel articulation where the voice were recorded simultaneously, which were provided from ATR-Promotions company. Here, MR image is analyzed based on the method of Yang et al. [8]. The procedure is following below.

1. Noise reduction of the vocal tract image by median filter.
2. Binarization process based on the average density of vocal tract boundary surface as threshold.
3. Detection of an edge from the smoothed image.
4. Extraction of the coordinate data of the contour of Tongue from the palate.

We have analyzed the MR images of 12 subjects to extract their tongue shapes, and have the result that Japanese vowel /u/ belongs to either front tongue or back tongue. This is the same as the estimation based on the reflections coefficients.

For example, we show respectively the tongue shape extracted from the MR image of a subject, and the calculated tongue shape from vocal tract area function in Fig. 6 and Fig. 7. From Fig. 6, we can see the follows.

1. During the articulation of /a/ and /o/, the whole tongue is lowered behind in the oral cavity and...
Karigome and Hanazaki: Use of Reflection Coefficients of Burg’s Method for Improvement of Visual Support Way in Pronunciation Practice

225

(a) After smoothing (b) Binary image

Fig. 5 Tongue shape estimated from MR image

(c) Separation of the tongue shape and palate

Fig. 6 Tongue shape estimated from MR image for Japanese vowels /a/, /i/, /u/, /e/, and /o/

(2) During the articulation of /i/, /e/, and /u/, the front tongue rises and front tongue is higher than the back tongue.

(3) The highest point of the tongue for /i/, /e/, and /u/ is higher than for /a/ and /o/.

Comparing the results in Fig. 6 and Fig. 7, the tongue shape estimated from the vocal tract area function is similar to the tongue shape extracted from MR image. Fig. 8 is showing that two clusters can be obtained taking high threshold of the distance. One denotes front tongue and the other does back tongue.

Table 4 Position of the articulation point of each subject

<table>
<thead>
<tr>
<th>subjects No.</th>
<th>Position of the articulation point from MR image</th>
<th>Position of the articulation point from voice data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>front</td>
<td>front</td>
</tr>
<tr>
<td>2</td>
<td>front</td>
<td>front</td>
</tr>
<tr>
<td>3</td>
<td>back</td>
<td>back</td>
</tr>
<tr>
<td>4</td>
<td>front</td>
<td>front</td>
</tr>
<tr>
<td>5</td>
<td>back</td>
<td>back</td>
</tr>
<tr>
<td>6</td>
<td>front</td>
<td>front</td>
</tr>
<tr>
<td>7</td>
<td>front</td>
<td>back</td>
</tr>
<tr>
<td>8</td>
<td>front</td>
<td>back</td>
</tr>
<tr>
<td>9</td>
<td>back</td>
<td>back</td>
</tr>
<tr>
<td>10</td>
<td>front</td>
<td>front</td>
</tr>
<tr>
<td>11</td>
<td>front</td>
<td>front</td>
</tr>
<tr>
<td>12</td>
<td>front</td>
<td>front</td>
</tr>
</tbody>
</table>
The 2nd formant is generally said to represent the front or back position of the high point of tongue. But the 2nd formant of /u/ is close to the formant of /a/ in Fig. 9, where the formants of the subject is shown. The high position of tongue of /u/ is confirmed to be close to /i/ and /u/ in Fig. 6, and this is confirmed in the estimation from the vocal tract area function.

In Table 4, the comparing with the highest positions of the tongue extracted from 12 MR images and the vocal tract area is shown. Japanese vowels /u/ of 9 subjects were judged to be belonging to the front tongue, though /u/ is said to belong to the back tongue by the formants. According to the MR images, variation of the highest position of the tongue might be depended on the manner of articulatory, especially the shape of the lip. During articulation of the vowel /u/, some subjects uttered protruding the lip.

Then we are going to investigating the relationship between the degree of protruding the lip and the highest position of the tongue. The distance between the position of the upper lip nodal part during the rest state and the uttering /u/ is adopted to represent the degree of protruding the lip. This distance is illustrated in Fig. 10. In Fig. 11, we are showing the degree of protruding the lip vs. the highest position of the tongue. In Table 5, their mean and their variance corresponding to the front and the back tongue are shown. These results suggest that the highest position tend to the front tongue as uttering with the lip protruding.

The instruction about the tongue shape is important in pronunciation training. In practice, teachers
who correct the pronunciation give the instructions such as “Please utter from the back of the throat”, “Please be uttered as clogged” and so on. From the above result, we obtained the suggestion that lip protrusion should be taken into consideration when generating pronunciation teaching. So the tongue shape estimation from the vocal tract area function may lead to development of effective support way.

7. Conclusions

We have investigated relationship between the acoustic feature or articulation manner of the speech sound and the reflection coefficients which are calculated in the AR model analysis of speech signal. According to principal component analysis for the reflection coefficients of Japanese 5 vowels, it was shown that the 1st principal component relates closely to the 2nd formant and specifies the front or back position of the highest point of the tongue. In addition, the vocal tract area function calculated from the reflection coefficients were analyzed by PCA and the tongue shape showing the highest point clearly was estimated from the 1st principal component. Comparing with the estimated tongue shape and those extracted from MR image taken during uttering the vowels, we confirmed that the discriminates between the front and the back vowel by the estimated tongue shape is more precise than by the formants.

As the results of analysis of MR image, we found out that the highest point of the tongue uttering Japanese vowel /u/ moves to front from back depending on the lip protruding, though this vowel is belong to the back vowel based on the formants. Even in this case, it was confirmed the estimated tongue shape specifies correctly the highest position of the tongue. This result showed that the reflection coefficient is able to display articulation state considering the lips protrusion and gave possibility to improve visual support way based on the formants.

Acknowledgements

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.

This work was supported by JSPS KAKENHI(C) Grant Number 16K02897

References


Authors

Ryosuke Karigome

Ryosuke Karigome was born on 2nd July, 1993. He received B.E. degree from Tokyo Denki University in 2016. Presently, he is a graduate student of of Science and Technology for Future Life in Tokyo Denki University. His research field include construction of pronunciation training system using speech signal processing and modeling.

Izumi Hanazaki (Member)

She received D. E. degree from Waseda University, Japan, in March 1986. She is currently a professor in Department of robotics and Mechatronics, School of Science and Technology for Future Life, in Tokyo Denki University. Her research interests include on time series analysis, speech signal analysis and stochastic system analysis.