Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 227-233 (2017)

www.astesj.com

Special Issue on Computer Systems, Information Technology, Electrical and Electronics Engineering

ASTESJ

ISSN: 2415-6698

How Effective is Using Lip Movement for Japanese Utterance Training

Miyuki Suganuma*[1], Tomoki Yamamura[2], Yuko Hoshino[2], Mitsuho Yamada[1]

[1]*Graduate school of Information and Telecommunication Engineering, Tokai University, 108-8619, Japan*

[2]*School of Information and Telecommunication Engineering, Tokai University, 108-8619, Japan*

ARTICLE INFO

Article history:
Received: 20 December, 2016
Accepted: 20 January, 2017
Online: 28 January, 2017

Keywords :
Lip movements
utterance
Japanese

ABSTRACT

Lip movements have long been the subject of research. There are many methods of lip movement recognition, such as the calculation of the amount of movement compared to a matching face pattern. In a previous study, we investigated utterance recognition based on the power spectrum by focusing on lip movements, which is one aspect of multimodal voice recognition systems. However, we found that the utterance recognition rate varied depending on the participants throughout our research. For this reason, we propose a training method for the purpose of utterance improvement in Japanese.

1. Introduction

In recent years, the rapid advancement of computer technologies has led to the widespread use of personal information which, when converted into digital data, can be read and changed without place and time restrictions. Therefore, large-scale processing of data is optimal. However, to use paper-medium information as digital information, it must be input by hand using a keyboard. As a way to simplify this work and make it more intuitive, voice recognition is an attractive technology. Data entry using voice recognition is easier to use than keyboard entry because it can be used hands-free. For this reason, many researchers have been researched voice recognition, for example, the non-contact type interface of command input and the individual identification [1,2]. On the other hand, voice recognition technology is not without problems, and the recognition rate can decline when it is used in noisy locations. Therefore, multimodal recognition has been investigated, because it can make use of both voice recognition and face recognition [3,4]. In our research, we have been studied focusing on lip movement recognition [5]. However, we found that the recognition rate of lip movement varies depending on the person who is speaking. In response to this, we are developing ways to improve participants' utterances so that they may be more easily recognized [6]. The purpose of our research is to make participants' utterances more fluent. For the utterance learner, improving their utterance is a great incentive to study Japanese utterance. With regard to studies to improve utterances, the researches that is focused on the tongue movement in vowel utterances not only Japanese also in other language such as English have been conducted [7]. Wilson measured the tongue movement by filling the lower jaw with an ultrasound sensor [7]. Electromyogram is also possible to measure the tongue movement [8], however, it is difficult to specify myoelectric signal along mixing other myoelectric signals with lip movements. Therefore, we propose the utterance training method only using acquired lip image while uttering, expanding the research of our lip feature point collecting application by lip movement because we considered the noninvasive utterance training can be conducted. Using simple training application that we developed in our laboratory enable to do self-utterance learning. Utterance learners can study by themselves until to be satisfied with the result of utterance training. Moreover, we also aim at make it possible for speech- and hearing-impaired individuals to use our lip feature point collection application by themselves.

This paper is an extension of work originally presented in ICCSE 2016 (The 11th international conference on computer science and education) [9]. We explain about speech training system which is developed in our laboratory in second section.

*Corresponding Author Name: Miyuki Suganuma, Graduate school of Information and Telecommunication Engineering, Tokai University, +810334411171, m.suganuma@hope.tokai-u.jp

Third section is method. Fourth section is results and discussions. Finally, we describe conclusion in fifth section.

2. Speech training system

In this research, we used a lip feature point collecting application that was developed in our laboratory [10]. This application was developed with Visual Studio 2010 as a Windows form application in the C++ language. FaceAPI, which is facial recognition software made by SeeingMachines in Australia, was used for the face recognition, as it is able to perform at high speeds and with high precision. The processing speed is estimated to be 0.3 seconds when the head does not move and a Core-2 Duo 2.4 GHz processor is used. Our equipment uses an i7 1.9 GHz processor and has sufficient processing speed for lip movement detection.

Figure 1 is an example of the display used for the acquisition of lip feature points. Figure 2 shows the display for lip movement training. When the subject is recognized by the lip feature point collecting application, the acquired face feature points can then be shown for training. To operate this application, a file for training must first be selected from the “Set up” menu and the file name typed to save data. Next, training is initiated by clicking on the “Start” button. When the subject closes his or her mouth, the application recognizes that person's mouth and the subject can start pronouncing words. If the subject is not recognized by the application, the “Re-recognition” button can be used. The “Stop” button stops recording. The lip movement training display is shown after recording is stopped. The red line on the lip movement training display (Figure. 2) shows the teacher's lip movements, which are used as model data; the black line shows the participant's lip movements. The two sets of lip movements can be compared by clicking on the “Start” button (Figure. 2).

3. Method

The present study was carried out to improve the learner’s Japanese utterance like veteran announcer's utterance by using a lip feature point collecting application that was developed in our laboratory.

The flow of lip movement training is shown in Figure 3. First, a file for training selects from the “Set up” menu. Second, the participant pronounces each sentence with the application by clicking on the “Start” button. Third, the participant compares the announcers’ lip movement data with their own as shown in Figure 2 as the black line. The participant can watch his or her lip movement and the teacher's lip movement any number of times by clicking “Start” in Figure 2. Nevertheless, the participant cannot see a line graph of lip movement. This flow is a set of the training. The training consisted of a total of 10 repetitions. Maximum two repetitions of the training were carried out per day in order to the participants not grow tired.

A model database for Japanese utterances was created in our previous research[7]. The model data were acquired from a television announcer whose lip motion was considered to be correct. The announcers were two veteran announcers, a man and a woman, from the NHK (Japan Broadcasting Corporation) Communications Training Institute. Based on the model data that we created, we trained 4 students from our university (three males and one female: aged 21 to 22). To improve their utterances, we used the lip feature point collecting application that is explained in the preceding section. The experiment was conducted in a studio at our university, and we used a Windows note PC (SONY VAIO PCG-81314N). The web camera screen resolution was 640x480. The set-up for the acquisition of utterance data is shown in Table 1. Japanese data were extracted from the book "Easy training for a good voice" [11], which includes 69 sentences from *"a"* to *"wa"* In order to acquire both long and short sentences, we used the "dandelion” text from a third-grade textbook of the Japanese language that is currently used for announcers’ utterance practice [12].

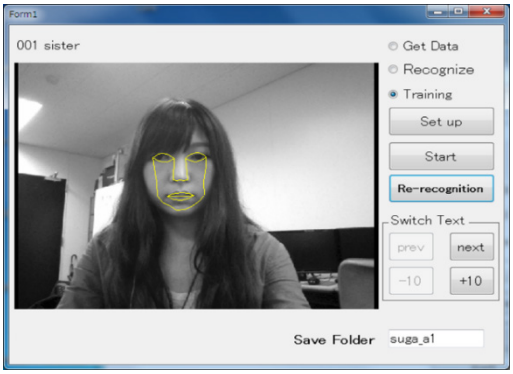

Figure 1. Display for the acquisition of lip future points

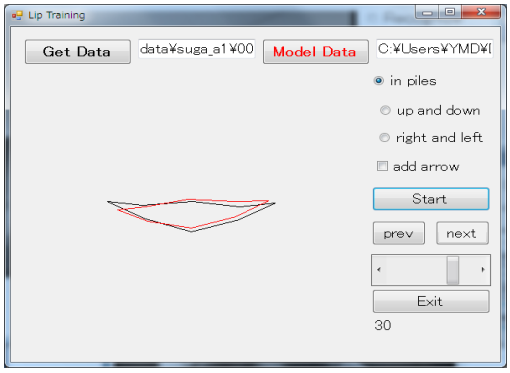

Figure 2. Display for lip movement training

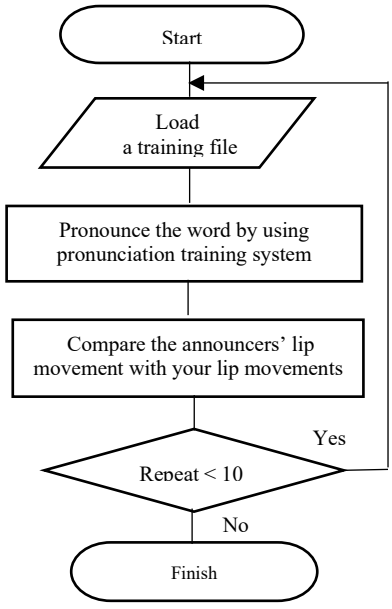

Figure 3. Flow chart for lip movement training

Table 1. Japanese data used

Attara aisouyoku aisatsu shinasai
Akogare-no aite-ni au
Ikigai-wo motomete ikou
Ima ijyou-no omoi-wo ireru
Uta-wo utatte usabarashi
Ukatsu-ni umai uso
Eiyo-yo eikou-yo eien-nare
Erai ekakisan-ga eranda-e
Oishii okashi-wo osusowake
Ookami-no ookina tooboe

4. Results and Discussion

4.1. Lip movement data

We carried out Japanese utterance training using the veteran announcers' model data. The participants were students from our university. We compared the model data and the participants' data to determine their differences. In this section, we show only the results for *"Ookami-no ookina tooboe"* as an example. Figure 4 shows the 5 lip feature points: upper lip, lower lip, left lip, right lip and chin that we collected for the utterance training. Generally speaking, human beings' mouth moves vertically symmetry and lower lip opens more widely than upper lip, so that we used lower lip movement and left lip movement for discussing lip movement.

Figure 5 shows the veteran announcer's lip movements. Figures 6 and 7 show the participants' lip movements in several repetitions of the training (1 repetition, 5 repetitions, 10 repetitions). The vertical axis indicates the amount of displacement (pixel). The screen resolution is 640x480, therefore actual lip opening displacement is about 0.5 cm if an amount of displacement of mouth is 10 pixel. The horizontal axis represents time. The blue line shows the left lip feature points. The red line shows the lower lip feature points.

In the 1st repetition of the training, none of the participants opened their mouths as widely as the veteran announcer. However, in the 5th and 10th repetitions of the training, their utterances became clearer little by little, because the amplitudes of their lip movements became large and they pronounced the sentences precisely by every chunk, such as *"Ookami-no" "ookina" "tooboe"*.

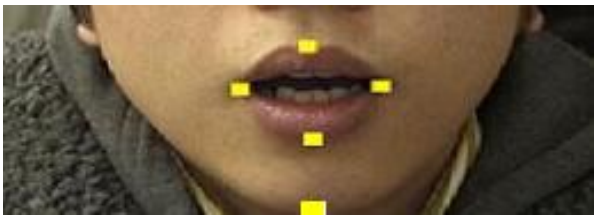

Figure 4. 5 lip feature points for the utterance training
(Upper lip, lower lip, left lip, right lip and chin)

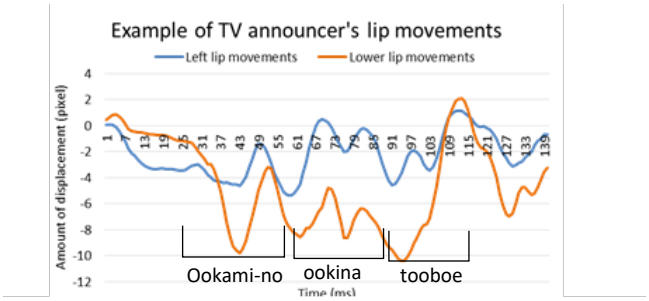

Figure 5. Veteran announcer's lip movement
"Ookami-no ookina toboe"

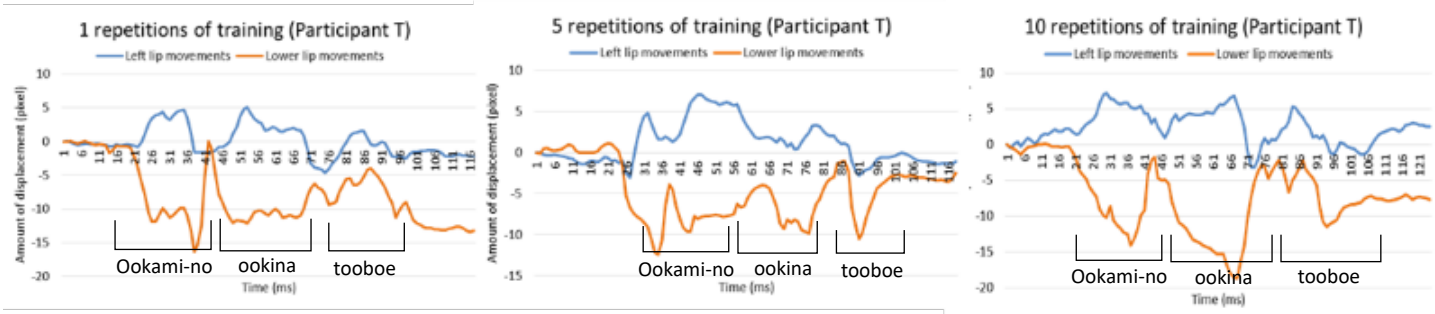

Figure 6. Participant T's lip movement *"Ookami-no ookina toboe"*

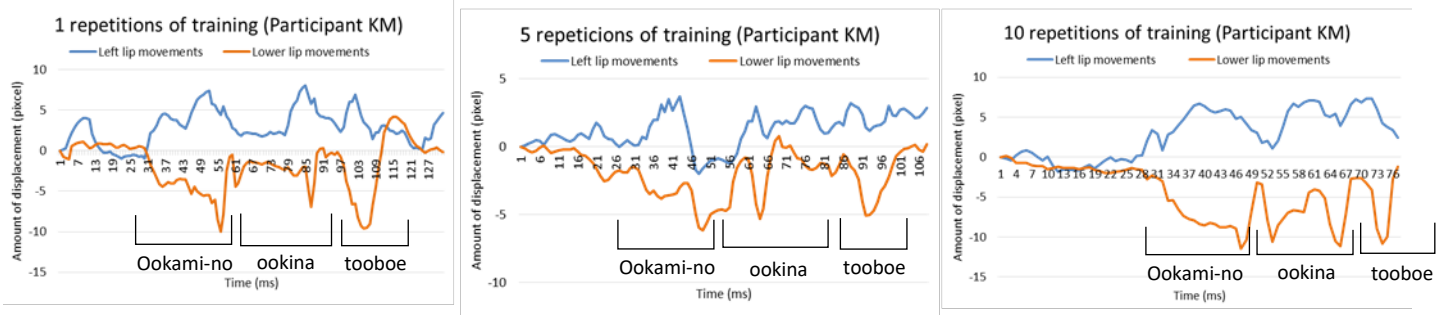

Figure 7. Participant KM's lip movement *"Ookami-no ookina toboe"*

4.2. Voice data

During the training, voice data were also recorded at the same time. Here, we show the voice data results. We used the paired comparison method to evaluate the voice data objectively.

The evaluators were students from our university including the participants who participated in the Japanese utterance training. There were 10 evaluators aged 20 to 22 years old. Participants did not evaluate their own voice data.

4.2.1. Pairwise comparison

Figure 8 shows an example of paired comparisons. We randomly chose the voice data from the 1st, 3rd, 5th, 7th and 10th repetitions of the training and evaluated these data by the paired comparison method. In this method, evaluators alternatively chose which voice is good by focusing on following items: articulation, speed and volume. These three items were uniquely chosen based on the veteran announcers' characteristics which can be able to improve their utterance [13]. In the articulation, evaluators compared sentences whether each sentences uttered clearly by every chunk such as, *"Attara" "aisouyoku" "aisatsu" "shinasai"*. In the voice speed, evaluators chose which sentence is easy to hear, also pronounced appropriate speed. In the voice volume, evaluators genuinely chose the sentence which is louder. At this time, voice volume was set by the evaluators before the training, and we instructed them not to change the volume during the training. These comparison items were explained to evaluators before training. We explained to the evaluators what items they will evaluate before the voice comparison was conducted. Each evaluator evaluated 4 repetitions of voice comparisons per participant. The total number of comparisons was 16.

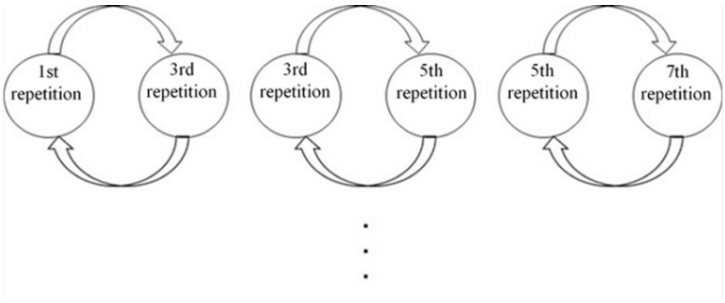

Figure 8. Example of paired comparison

We show the two example sentences. Figure 9 shows the results for *"Attara aisouyoku aisatsu shinasai"*. It shows the average of all comparison results and the standard deviation. First line indicates the articulation. Second line is the voice speed. Third line is the voice volume. Likewise, Figure 10 is the results for *"Ikigai-wo motomete ikou"*. Table 2 shows the summarized results of each evaluation items improvement and the result of t-test in the sentence of *"Attara aisouyoku aisatsu shinasai"*. Table 3 shows the summarized results of each evaluation items improvement and the result of t-test in the sentence of *"Ikigai-wo motomete ikou"* same as Table 2.

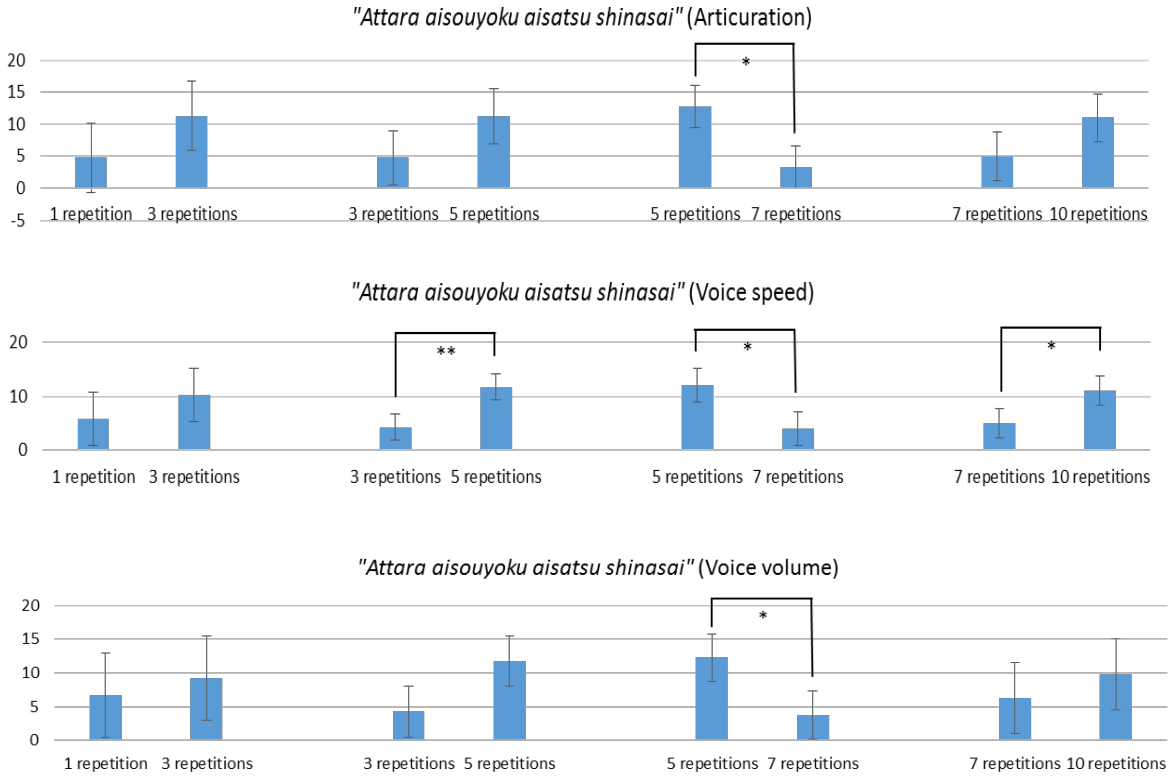

Figure 9. Results of voice comparison *"Attara aisouyoku aisatsu shinasai"*

Table 2. Summarized results of each evaluation items improvement and the result of t-test
"Attara aisouyoku aisatsu shinasai".

Attra aisouyoku aisatsu shinasai						
Set of training	Articulation		Voice speed		Voice volume	
	Improvement (up or down)	Significant differences	Improvement (up or down)	Significant differences	Improvement (up or down)	Significant differences
1-3 repetitions	up	n.s	up	n.s	up	n.s
3-5 repetitions	up	n.s	up	**	up	n.s
5-7 repetitions	down	*	down	*	down	*
7-10 repetitions	up	n.s	up	*	up	n.s

n.s.: non-significant, *: p<0.05, ** p<0.01

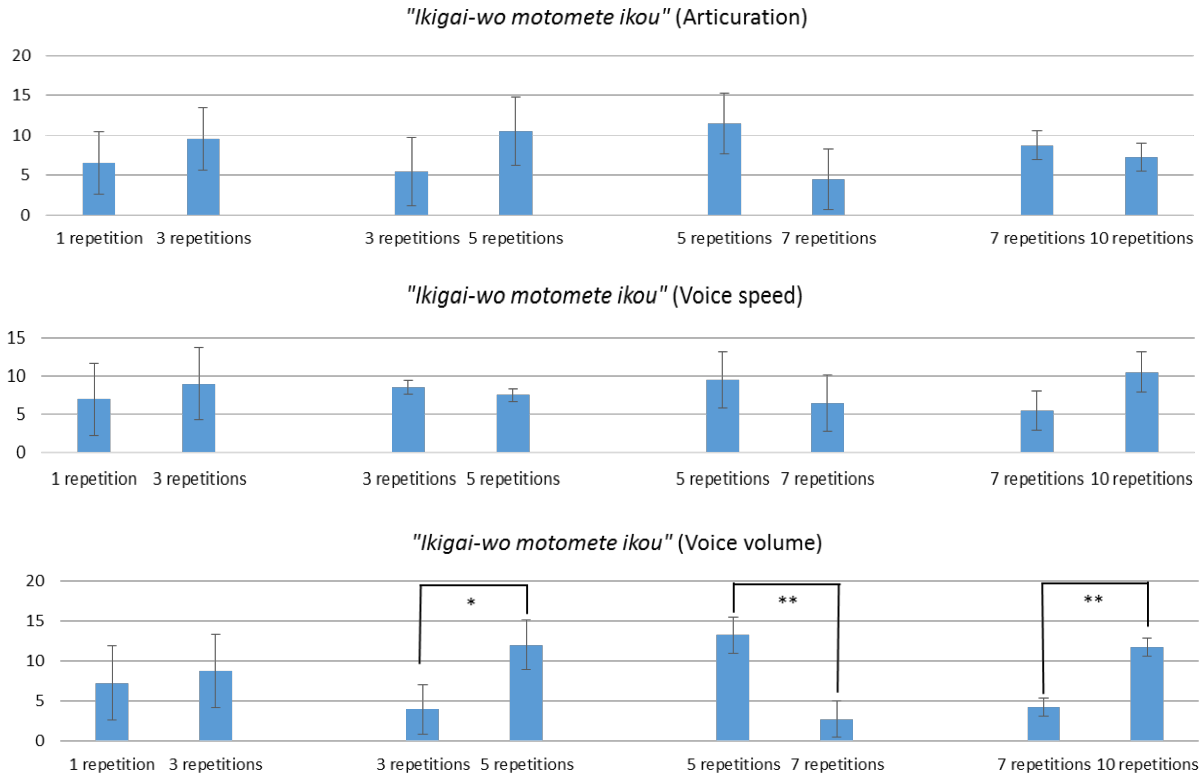

Figure 10. Results of voice comparison *"Ikigai-wo motomete ikou"*

Table 3. Summarized results of each evaluation items improvement and the result of t-test
"Ikigai-wo motomete ikou".

Ikigai-wo motomete ikou						
Set of training	Articulation		Voice speed		Voice volume	
	Improvement (up or down)	Significant differences	Improvement (up or down)	Significant differences	Improvement (up or down)	Significant differences
1-3 repetitions	up	n.s	up	n.s	up	n.s
3-5 repetitions	up	n.s	down	n.s	up	*
5-7 repetitions	down	n.s	down	n.s	down	**
7-10 repetitions	down	n.s	up	n.s	up	**

n.s.: non-significant, *: p<0.05, ** p<0.01

Throughout 10 repetitions of training, transitions of improvement were small from 1st repetition to 3rd repetitions, then the transitions became large from 3rd repetitions to 5th repetitions. However, sometimes improvements decreased from 5 repetitions of training to 7 repetitions of training. After that, improvement tendency became large from 7th repetitions to 10th repetitions. We could see that utterance improvement is not seen in each repetitions of training. Nevertheless, improvement tendency is seen after 5 to 7 repetitions of training because participants came out of a slump. This tendency is in common between all vowels and it is considered to be one of the learning characteristics.

In addition, we carried out an unpaired two-tailed t-test regarding the average of the marking results. The t-test result of paired comparison shows in Figures 9 and 10 (also Tables 2 and 3) in "*" or "**" if there are significant differences. There was significant difference between 3 repetitions of the training and 5 repetitions of the training in voice speed item sentence *"Attara aisouyoku aisatsu shinasai"* ($p=0.008<0.05$) in Figure 9. There were significant difference between 5 repetitions of the training and 7 repetitions of the training, and between 7 repetitions of the training and 10 repetitions of the training in voice volume item sentence *"Ikigai-wo motomete ikou"* (both $p=0.008<0.01$) in Figure 10. All vowel's utterance are improved in all 3 evaluate items from 3rd repetitions to 5th repetitions and from 7th repetitions to 10th repetitions, even though some exception is seen. We also carried out an unpaired two-tailed t-test in other sentences. There were significant differences in /a/, /i/ and /o/ sounds. Whereas, there were no significant differences in /u/ and /e/ sound. When the native Japanese speakers utter /a/, /i/ and /o/ sounds, they usually open their mouth widely so that we consider significant differences could see. From these results, we confirmed that the more training the participants had, the better the evaluations they received.

4.2.2. Formant analysis

We used formant analysis in addition to pairwise comparison method because vowels are commonly expressed in the International Phonetic Association (IPA) vowel chart in English-speaking country. The IPA vowel chart shows the position of the tongue and how the mouth is opened for each sound. If the frequency of the 1st formant is high, the tongue is positioned in the front of the mouth. If the frequency of the 2nd formant is high, the position of the tongue in the mouth is low. [14]. Figure 11 shows average Japanese men's vowel formant. Mori states that the 1st and 2nd formants can classify vowels regardless of race and gender [15]. Otsuka points out that the sound spectrograms of the 1st and 2nd formants are helpful in teaching the pronunciation of English vowels [16]. We have been also researched English pronunciation training using application which was used in this research. We used this evaluation method and we could made close to learners' pronunciation to native English speakers' pronunciation [17].

In this paper, we show the scatter diagram of the formants of the participants about *"Eiyo-yo eikou-yo eien-nare"* from 1 repetition to 10 repetitions of the training as an example in Figure 12 because this sentence includes many /e/ sound. The circle marker indicates 1 repetition of the training. The square marker is 5 repetitions of the training. The diamond marker is 10 repetitions of the training. We can see that each markers gathered throughout the training. Moreover, in compared to the average vowel formant diagram in Figure 12 and the result of the sentence *"Eiyo-yo eikou-yo eien-nare"*, we can see that their formants are significantly close. When the native Japanese speaker move our mouth from *"e"* to *"i"*, they usually move their mouth from lengthwise direction to sideways direction. Therefore, we consider that participants could pay attention to utter the sentence clearly.

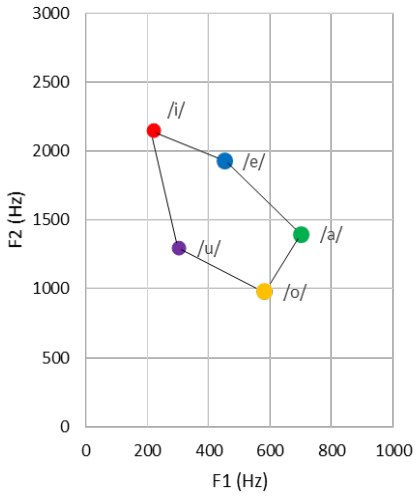

Figure 11. Average Japanese men's vowel formant

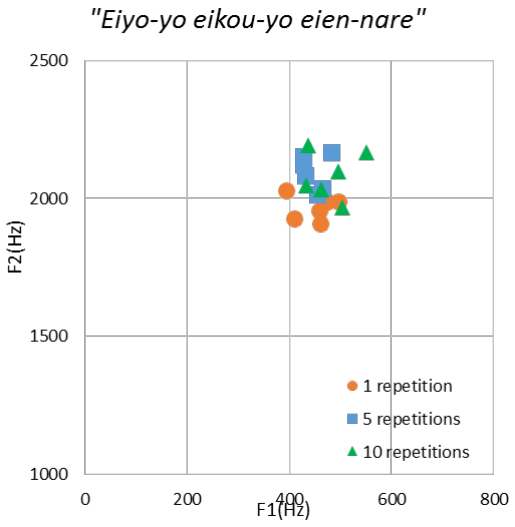

Figure 12. Scatter diagram of the participants about each repetitions of training *"Eiyo-yo eikou-yo eien-nare"*

5. Conclusion

In this research, we evaluated the usefulness of Japanese utterance training using the lip feature point collecting application in order to improve people's utterance ability and prevent a decline in the voice recognition rate. To train participants in making utterances, we used veteran announcers' data as the model data.

We evaluated not only the lip movement results but also the voice data results using the pairwise comparison method and formant analysis. As for the lip movement results, we confirmed that the participants' utterances improved in comparison to the first repetition of training, because the amplitudes of their lip movement's became larger. Also participants' could be able to pronounce clearly by each chunk such as *"Ookami-no" "ookina" "tooboe"*. With regard to the voice comparisons, the evaluation showed that the training seemed have a good effect up through ten repetitions of the training, especially /a/, /i/ and /o/ sounds. Whereas, we could not get a training effect by each repetitions of training because there are training slump or fatigue from 5 repetitions of training to 7 repetitions of training. However, this result indicated that this is one of the learning characteristics. We need to find how to make participants not bored from 5 repetitions of training to 7 repetitions of training to solve this problem. We considered that the more training the participants had the better the evaluations they received, if we can find the solution. In addition, some remarkable improvement is seen in the result of formant analysis, for example, *"Eiyo-yo eikou-yo eien-nare"* because Japanese vowel of /e/ sound in the sentence grow to the Japanese average /e/ sound. These results suggest that Japanese utterance training using the lip feature point collecting application is useful.

In the future, we need let to use the lip feature point collecting application to more participants' and wide range of age groups. We plan to analyze the lip movement data more in detail in order to enhance the reliability of outcomes, for example, an open area of mouth, a count of amplitudes and a calculation of utterance time. Also, we would like to try voice comparison by expanding the sample differences as a future task because our sample differences were small such as 1 repetitions and 3 repetitions of the training. Moreover, we think the relationships between smooth graph and voice quality is also necessary to consider.

Acknowledgment

We are thankful to the television announcers with NHK communications training institute who cooperated with the creation of an utterance database. This work was supported by JSPS KAKENHI Grant Number (25330418) and (16K01566).

References

[1] Y. Sato, Y. Kageyama and M. Nishida, "Proposal of Non-Contact Type Interface of Command Input Using Lip Motion Features" The transactions of the Institute of Electrical Engineers of Japan. C, A publication of Electronics, Information and System Society 129(10), 1865-1873, 2009.

[2] Y. Shirasawa, S. Miura, M. Nishida, Y. Kageyama and S. Kurisu, "Method for Identifying Individuals Using Lip Motion Features" Journal of Image Information and Television Engineering, 60(12), 1964-1970, 2006.

[3] M. Yoshikawa, T. Shinozaki, K Iwano and S. Furui, "Multimodal Speech Recognition Based on Lightweight Visual Features" IEICE Information and Systems, J95-D (3), 618-627, 2012.

[4] S. Hayamizu and T. Takezawa, "Trends in Research on Multimodal Information Integration System" Jornal of the Japanese Society for Artificial Intelligence, 13(2), 206-211, 1998.

[5] E. Wakamatsu, Y. Hoshino and M. Yamada, "Proposal for an utterance training method based on lip movements" in IMQA 2014 (The Seventh International Workshop on Image Media Quality and its Applications), 2014.

[6] M. Suganuma, T. Yamamura, Y. Hoshino and M. Yamada, "How to Evaluate English Pronunciation Learning by Lip Movements" in IMQA 2016, 2016.

[7] I. Wilson, "Using ultrasound for teaching and researching articulation" Acoust. Sci. & Tech., 35 (6), 285-289, 2014.

[8] C. Huang, C. Chen and H. Chung, "The Review of Applications and Measurements in Facial Electrography" Journal of Medical and Biological Engineering, 25(1), 15-20, 2004.

[9] M. Suganuma, T. Yamamura, Y. Hoshino and M. Yamada, "Effect of Japanese utterance training using lip movement" in ICCSE 2016 (The 11th international conference on computer science and education), 2016.

[10] E. Wakamatsu, K. Kikuchi and M. Yamada, "Development and Measurement of an Automatic Word Recognition by Lip Movements" in IMQA 2013, 2013.

[11] E. Fukushima, Easy training for a good voice, Seibido Printing, 2006, in Japanese.

[12] Dandelion, New version of the Japanese elementary school 2nd year textbook, Tokyo Books, 1985, in Japanese.

[13] S. Suzuike, "The study on the significance of the non-verbal elements of the human voice in the communication," Bachelor thesis , Waseda University, 2010.

[14] S. Imaizumi, Phonetics and Linguistics for speech-language-hearing therapist, Igaku -Shoin Ltd., 2009.

[15] T. Mori, "Multilingual vowel sound comparison analysis by the formant in sound, the articulatory phonetics," Bachelor thesis, Nagoya University, 2015.

[16] S. Otsuka, "A Trial in Utilizing Formant Values for Teaching Vowel Pronunciation," Science reports of Tokyo Woman's Christian University, 2014.

[17] M. Suganuma, T. Yamamura, Y. Hoshino and M. Yamada "Proposal of the Way of English Pronunciation Training Evaluation by Lip Movement" Journal of Japan Personal Computer Application Technology Society, 11(2), 2017. (will be published)