IEICE TRANS. INF. & SYST., VOL.E102–D, NO.1 JANUARY 2019

PAPER

Visual Emphasis of Lip Protrusion for Pronunciation Learning

Siyang YU[†a)], ***Nonmember***, **Kazuaki KONDO**[††], **Yuichi NAKAMURA**[††], ***Members***, **Takayuki NAKAJIMA**[†††], ***Nonmember***, **Hiroaki NANJO**[††], ***Member***, ***and*** **Masatake DANTSUJI**[††], ***Nonmember***

SUMMARY Pronunciation is a fundamental factor in speaking and listening. However, instructions for important articulation have not been sufficiently provided in conventional computer-assisted language learning (CALL) systems. One typical case is the articulation of rounded vowels. Although lip protrusion is essential for their correct pronunciation, the perception of lip protrusion is often difficult for beginners. To tackle this issue, we propose an innovative method that will provide a comprehensive visual explanation for articulation. Lip movements are three-dimensionally measured, and face images or videos are pseudocoloured on the basis of the movements. The coloured regions represent the lip protrusion of rounded vowels. To verify the learning effect of the proposed method, we conducted experiments with Japanese undergraduates in Chinese classes. The results showed that our method has advantages over conventional video materials.

key words: *lip protrusion, pseudocolouring, rounded vowel, visual support in CALL*

1. Introduction

In the field of second language (L2) education, instructors and researchers try to improve the four language skills of students, i.e. speaking, listening, reading and writing [1]–[3] in many different ways. Pronunciation is a fundamental factor in speaking and listening. Intelligible pronunciation not only enables students to understand and be understood but also helps them to monitor their speech on the basis of input from the environment. Moreover, it has a great impact on their confidence to communicate in an engaging manner [4].

Correct pronunciation requires the accurate position of various articulators, such as the tongue, lips and jaw. To make clear and distinct word sounds, different speech organs work together to create diverse obstructions and/or oral cavities so as to shape the air in a particular fashion as it goes from the inside to the outside of the body.

Some evidence supports the importance of visual perception for correct pronunciation. Even with regard to native speakers, sighted people can produce vowels more distinguishably and precisely than congenitally blind people can, according to research in [5]. They also found that due to visual deprivation, congenitally blind adults use more tongue variations in the implementation of vowel contrasts; however, this does not completely compensate for the reduced degree of upper lip protrusion in those native speakers. Another piece of evidence is the effect of visual information in speech perception. Our perceived pronunciation is the result of the interaction between hearing and vision. The McGurk effect [6] is a typical phenomenon that demonstrates such a symbiotic outcome in speech perception. Inconsistency between the auditory component and the visual component in terms of pronunciation can lead to illusion, i.e. the perception of a third sound.

In language education, computer-assisted language learning (CALL) has advanced from early one-way instruction to being able to provide feedback to students regarding their pronunciation [3]. According to [2], the simplest application just works as a digital recorder; learners can record their own pronunciations for comparison with a native speaker's. A step further than a digital recorder is the use of speech visualisation, for instance, waveforms, spectrograms, formant frequencies, pitches, contours and so on, as seen in [7]. A more sophisticated application can be achieved by employing automatic speech recognition (ASR). Tsubota et al [8] developed a system that estimates the intelligibility of students' speech and ranks their errors. Practice exercises for improvement, as well as instructions for correcting errors, are provided afterwards.

Most researchers and practitioners using CALL systems have paid considerable attention to audio, whereas some important aspects of articulation have not been well explored. We have seen that Japanese learners, particularly beginners, experience difficulty in learning some Chinese pronunciations even with the support of a CALL system. We focused on this problem and investigated a method that would provide a more comprehensive visual explanation of articulation for students.

In this paper, our objective is explained in Chapter 2, i.e. the problems in learning pronunciation and our ideas to resolve the issues. Details of the system that we designed and its implementation are explained in Chapters 3 and 4, respectively. In Chapter 5, the experimental results of verifying the effects of our method are presented.

2. Problem and Objective of Research

Different languages do not have the same phonetic sys-

Manuscript received December 21, 2017.
Manuscript revised July 27, 2018.
Manuscript publicized October 22, 2018.

[†]The author is with Graduate School of Engineering, Kyoto University, Kyoto-shi, 606–8501 Japan.
[††]The authors are with Academic Center for Computing and Media Studies, Kyoto University, Kyoto-shi, 606–8501 Japan.
[†††]The author is with Graduate school of Human and Environmental Studies, Kyoto University, Kyoto-shi, 606–8501 Japan.
a) E-mail: yusiyang@ccm.media.kyoto-u.ac.jp

DOI: 10.1587/transinf.2017EDP7411

tems. This diversity requires learners to use new movements of articulators for new pronunciations and also necessitates a lot of practice to achieve proficiency. As a typical example, Japanese learners often encounter difficulties in 'rounded pronunciations' that require lip protrusion lacking in Japanese pronunciation, since the Japanese do not have clear distinction between rounded vowels and unrounded vowels in daily conversation. In fact, Duan et al suggested that one of the dominant Chinese mispronunciation patterns by Japanese learners is "Lip rounded and spreading" [9] which indicates a need to learn those pronunciations. Other languages, such as German and French, also have clear distinction between rounded and unrounded vowels, and its learning is important for Japanese learners. For instance, '*yu*', that is, '*fish*' with a second tone in Chinese or '*über*', that is, '*over*' in German. The Chinese example has an additional difficulty because the Japanese language has a different pronunciation for the same romanised symbol. Japanese learners often try to produce a 'rounded' pronunciation with the mouth shape for 'unrounded' pronunciation for the same romanised symbol, and it makes an incorrect sound. For this reason, we focused on a visual aid that makes Japanese learners aware of the differences of mouth shapes of rounded and unrounded vowels.

For training in such pronunciations, explanations given by written texts are far from comprehensible, and even conventional multimedia approaches of showing pictures or videos do not give sufficient explanations about the articulations. We need a more comprehensible presentation of articulations because rounding has a three-dimensional shape deformation, and lip protrusion cannot be easily recognised through ordinary pictures or movies. Since current CALL systems do not provide enough functions for this purpose, we need to devise new functions.

In this study, we consider the following new functions of CALL systems:

(a) Demonstrate how native speakers pronounce words by showing articulation clearly and distinctly with three-dimensional information.

(b) Demonstrate how learners pronounce the words in the same way as above and enable them to check their correctness and/or weakness.

To achieve this, we obviously need to carry out a three-dimensional (3D) sensing of the face, particularly around the mouth. For this purpose, we use a Red-Green-Blue-Depth (RGB-D) camera. This type of camera has become inexpensive and can easily be connected to an ordinary computer.

Our scheme is as follows. Videos with 3D information are obtained using an RGB-D camera, and then image enhancement (pseudocolouring in our method) is applied to the captured images or videos. We expect that the camera and image processing software can be easily installed on the computers of both teachers and learners. On teachers' computers, pronunciations of native speakers are recorded as they are used as educational materials for learners to watch and learn. On the learners' computers, images and data are obtained in the same way, and they are used to check the correctness of their articulation.

To verify the arrangement, we conducted an experiment as described in (a). The method of visualisation is the same for (b); however, the learning effects are different. Therefore, we concentrated on the former and evaluated how learners could improve their pronunciation by watching the enhanced videos obtained by teachers. Our experimental results suggested that Japanese learners can improve their pronunciations once they are aware of mouth shapes.

3. System for Visualisation

3.1 Raw Data Acquisition

We employed Kinect v2 as an RGB-D sensor, which is ordinarily available at a low cost. The accuracy of Kinect v2 was investigated in [10]. The average depth error is less than 2 mm in a certain area; however, it often has worse measurement. It is reported that the measured depth value ranges from 1996 mm to 2004 mm for the true depth of 2000 mm. On the other hand, precision is much better as reported in [11] and [12]; average standard deviation is lower than 1.5 mm and remains stable during recording. For our purposes, precision is essential because protrusion can be measured as the relative depth change from the normal position. Lip protrusion for Chinese rounded pronunciation approximately ranges from 5 mm to 10 mm. In most cases, the precision is enough to detect it.

The quality of audio recording in Kinect v2, in contrast, is not satisfactory for language learning. For this purpose, we used another audio recording device, which is commonly used in audio recording. The process of obtaining pronunciation clips are as follows (Fig. 1). First, audio and video data are recorded with timestamps provided by the software development kit (SDK) of Kinect v2. They are segmented based on the values of audio data in which a silent period has significantly smaller values than pronunciation periods. Audio is recorded at the same time using the additional audio device. Second, audio data recorded by the additional audio device and the video data recorded by Kinect are synchronized by aligning the obtained data.

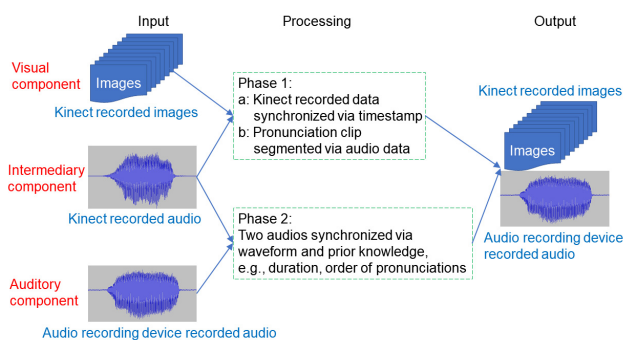

Fig. 1 Synchronization of video data taken by Kinect and audio data taken by audio recording device

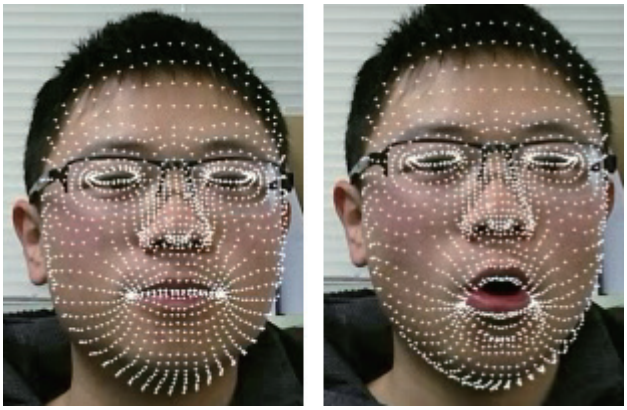

Fig. 2 Face and facial points detection result

3.2 Face and Feature Detection

More than 1000 facial points can be detected through deploying Kinect SDK. Figure 2 gives examples of detection results. We can use the results to locate the face and the mouth area in particular. However, they do not cover the mouth region seamlessly and points around the lips are influenced by mouth movements as shown on the right-hand side of Fig. 2. Therefore, another type of processing that emphasises lip articulation is necessary, details of which are explained in the next chapter.

In the SDK of Kinect v2, the face-tracking algorithm requires that the torso is also detected simultaneously. A speaker needs to sit at a certain distance from Kinect to ensure that the upper body is visible.

3.3 Reference Point and Protrusion Measurement

Lip protrusion is the relative depth change of the lip to the other parts of the face. To measure the amount of protrusion, we need a reference point that satisfies two conditions that are as follows: (1) it should be steadily observed regardless of ordinary body movements and (2) its location should not be affected by protrusion or any other kind of mouth movement. The tip of the nose satisfies these requirements well as a reference point. Therefore, the relative distance to the tip of the nose is calculated for each pixel around the lips and is used to judge protrusion.

First, the amount of depth difference at each point p_i is calculated as follows:

$$\Delta d = d_i - d_0$$

where d_i is the depth of the i-th pixel, p_0 is the reference point, the depth of which is d_0.

Then, for each point on or around the lips, if its depth difference is smaller than the predetermined threshold, we regard the point as being protruded.

4. Visual Enhancement by Pseudocolouring

Face deformation parallel to the image plane can easily be perceived using ordinary videos. In contrast, lip protrusion is perpendicular to the image plane and its perception is often difficult. To solve this problem, we use pseudocolouring on the basis of the measured depth of the lips.

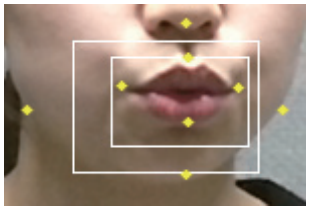

Fig. 3 Mouth area: lip subarea and non-lip subarea; yellow points are facial landmarks used for area determination, white rectangles are determined results

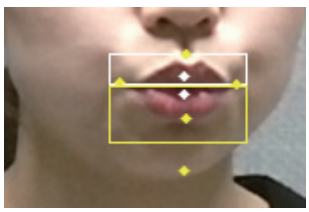

Fig. 4 Upper lip and lower lip subarea: yellow points are used for lip subarea segmentation as in Fig. 3; white points are used for further segmentation

4.1 Colourising Area Selection

We designed the system so that only a fixed area around the mouth is pseudocolourised because colour changes over a wide area of the face would make viewers feel uncomfortable. We, therefore, consider two different subareas, the lip subarea and the non-lip subarea, as illustrated in Fig. 3. For the lip subarea, it is coloured vividly with an attention-grabbing colour if the lips are protruded. For the non-lip subarea, it is coloured to provide a contrast if the lip subarea is protruded.

The lip subarea is further segmented into the upper lip subarea and the lower lip subarea as illustrated in Fig. 4. The physiological characteristic that means that the upper lip bulges slightly more in comparison to the lower lip sometimes results in unbalanced colourising.

4.2 Pseudocolouring Method

We designed the pseudocolouring on the basis of the Hue-Saturation-Intensity (HSI) colour space. Hue is used to represent the amount of depth change. Saturation is used to emphasise the important area of lip protrusion. Intensity is maintained as much as possible to give the original two-dimensional information as it is.

First, we convert RGB values to HSI values according to [13]. Details are given in the Appendix. Then, the HSI value is modified using the following method.

As we mentioned above, the intensity is kept the same as the input, i.e. $\tilde{I} = I$. Hue is determined according to

Table 1 Hue calculation

Mouth area level	Depth difference range Upper lip / lower lip (mm)	Hue value range (degree)
Significant protrusion	Upper lip subarea: $\Delta d \leq 20$ Lower lip subarea: $\Delta d \leq 27$	Red-based: [0, 20] [0, 27]
Low-protrusion or non-protrusion	Upper lip subarea: $20 < \Delta d \leq 60$ Lower lip subarea: $27 < \Delta d \leq 60$ Non-lip subarea: $\Delta d \leq 60$	Green-based: (120, 160] (120, 153] [120, 180]
Skin background	$60 < \Delta d \leq 100$	Blue-based: (240, 280]
Non-target area	$100 < \Delta d$	Original hue value

the relative depth difference to the reference point for each pixel. One important phenomenon that we noticed in our experiments is that if the hue changes continuously, it does not draw the viewers' attention very much. Significant discontinuity, i.e. coarse quantisation, is necessary for the changes to be noticed, i.e. protrusion. For this purpose, we quantise hue into three levels that correspond to three ranges of relative depth (see Table 1). Hue value is then modified via the formula shown in the Appendix. The upper lip and the lower lip usually have a slightly different amount of protrusion; hence, we designed slightly different corresponding ranges of distance between them. Note that the parameters concerning depth ranges are adjusted to the native speaker who provided pronunciation samples for our video-based materials.

Saturation is the most difficult part because not all the value combinations of hue, saturation and intensity can be converted into a valid range of RGB values. To cope with this problem, the maximum saturation value allowed with the values of hue and intensity that are modified in the above ways is calculated first. (The Appendix gives the actual method). The saturation value is determined on the basis of the maximum value and depth difference Δd for each hue range (shown in Table 1).

$$\tilde{S}_{lip} = \begin{cases} S_{max} & \text{hue is around red} \\ S_{max}\left(0.8 - (\Delta d - 20)/100\right) & \text{hue is around green: upper lip} \\ S_{max}\left(0.8 - (\Delta d - 27)/100\right) & \text{hue is around green: lower lip} \\ S_{max}\left(0.8 - (\Delta d - 60)/100\right) & \text{hue is around blue} \end{cases}$$

$$\tilde{S}_{non-lip} = \begin{cases} S_{max}\left(1.0 - (\Delta d/2)/100\right) & \text{hue is around green} \\ S_{max}\left(0.6 - (\Delta d - 60)/100\right) & \text{hue is around blue} \end{cases}$$

The maximum allowed saturation is assigned to a red-based hue of significant protrusion. The saturation value assigned to a green-based range of minor or no protrusion and that assigned to a blue-based range of skin background are determined as being proportional to the maximum saturation.

After the HSI values are determined using the above method, they are converted into RGB values as shown in the Appendix.

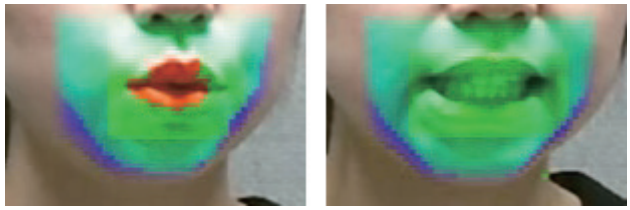

Fig. 5 Example of final result of pseudocolouring: rounded and unrounded

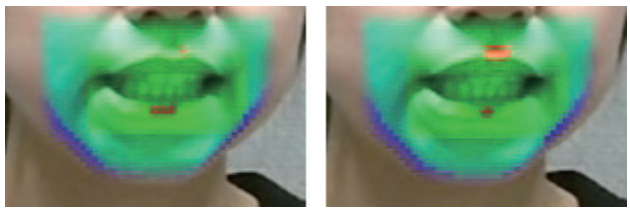

Fig. 6 Example of final result of pseudocolouring: noise included

4.3 Smoothing Around Borders

The discontinuity of color around borders can make images very unnatural and draws unnecessary attention. To avoid this effect, pseudocolours and original colours are blended proportionally on the basis of the distance from the border. Figure 5 shows an example of the final result of pseudocolouring. Figure 6 shows an example with non-negligible noise, and it is still acceptable.

5. Experiments

5.1 Objective

The experiments were conducted to confirm the effect of our proposed method on actual learners. We chose Chinese pronunciation as a target. We gathered 43 students from Kyoto University as participants: 10 were beginners who had not learned Chinese, 23 had learned Chinese for nearly one semester and 10 were students with experience of learning Chinese for approximately a year and a half.

To compare the effect on learning that our method has with the effect that conventional videos have, participants were randomly separated into two groups: Group A used conventional videos of a native speaker, whereas Group B used the native speaker's videos prepared using our method. We did not separate learners by learning periods because the statistics can be more reliable by the number of samples without separation. Conversely, it is not intended to evaluate the effects of learning periods.

Given that the two groups might have different levels of knowledge or ability as starting point, for better reliability of statistics, we used an indirect method in which advantages to the baseline methods described below are parameterised for both groups, and then the efficacy is compared on the basis

of the parameters. This idea is based on common methods in education, medicine, and others. For statistics in those fields, applying two or more different treatments to the same person is often avoided, e.g., different teaching methods, different medicine, etc. Otherwise, learning effects or curative effects would inevitably affect the statistics. Moreover, pre-test often affects the internal states of learners because they could be aware of what is focused upon and they could recall them if the pre-test is closely related to the content of statistics. This problem has been intensively examined, and not a few ideas and methods were proposed [14]. Rubin defined three categories of missing data [15], [16], based on which our experiments correspond to missing completely at random (MCAR) condition. If MCAR condition is satisfied, mean, regression, or some of other statistics can safely be compared between two groups.

5.2 Content

We chose a pair of unrounded and rounded vowels, /i/ and /y/ in the International Phonetic Alphabet (IPA) format. Both vowels are situated in the same position on the IPA chart; however, they have different degrees of lip protrusion. The corresponding formats in Chinese Pinyin notation are '*i*' and '*ü*'-symbolised as '*u*' when associates with '*j*', '*q*', '*x*' and '*y*', respectively.

They can be associated with six consonants-'*j*', '*l*', '*n*', '*q*', '*x*' and '*y*'-and four tones in Chinese. Consequently, we have 48 words in total. Each word is used 12 times for each participant to obtain sufficient samples for statistical analysis.

Four types of learning materials for those words were prepared: (a) pinyin symbol, (b) pinyin symbol with audio, (c) combination of pinyin symbol and common video (audio included) and (d) combination of pinyin symbol and pseudocolourised video (audio included). The 48 words are randomised into three groups, each group contains eight rounded words and eight unrounded words. One group is presented in pinyin material (a); another group is in audio material (b); the last group is in 2 types of material, both (c) and (d). Figure 7 gives the examples of (a) and (b). Figure 8 provides examples for (c) and (d). The left half demonstrates the examples of the mouth area used in the experiment; upper one is for (c), lower one is for (d). On the right is a full-face image of another native that gives an overall impression of video-based learning materials. The sequence of words is randomised every time to avoid an order effect. For (c) and (d), the pinyin symbol is placed near the mouth. This manoeuvre can keep participants' attention focused around the mouth. Participants in Group A perform the pronunciation for (a), (b) and (c); participants in Group B perform for (a), (b) and (d).

5.3 Scoring of Pronunciation

Every pronunciation made by the participants was evaluated by three Chinese native speakers who had experience

Fig. 7 Examples of pinyin symbol (a) and audio (b)

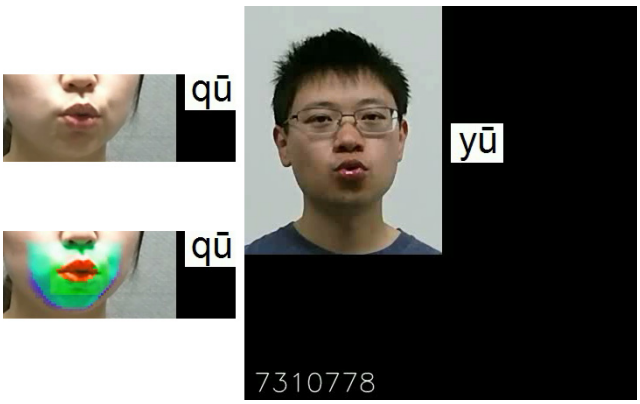

Fig. 8 Examples of common video (c) and pseudocoloured video (d)

in teaching the Chinese language to Japanese students. They judged whether the participants had made correct pronunciations by listening to audio files, with informed which pronunciations the participants were requested to pronounce. In particular, each pronunciation was evaluated as the correct pronunciation or another incorrect pronunciation. The final evaluation results were obtained on the basis of a majority vote, i.e. we accepted the results that two or more evaluators supported.

Arithmetic mean of Cohen's kappa [17], [18] of every pairwise native evaluators on rounded vowel, unrounded vowel, consonant and tone were calculated to evaluate inter-rater reliability (Table 2). According to the guidelines for interpreting kappa values [19], consonant and tone have almost 'perfect' agreement, rounded vowel has 'fair' agreement and unrounded vowel has 'moderate' agreement. This result suggests that learners' pronunciation of rounded and unrounded vowel is sometimes ambiguous for which evaluators do not have a good match. However, the value of kappa indicates 'fair' or 'moderate' based on the guideline, which is still acceptable. Therefore, we conclude that each evaluation can be trusted if majority of evaluators support them.

5.4 Results and Discussion

Since rounded vowels are our focus concerning articulation that requires lip protrusion, we first present the detailed re-

Table 2 Inter-rater reliability values

Pronunciation element:	Arithmetic mean of Cohen's kappa:
Consonant	0.89
Rounded Vowel	0.39
Unrounded Vowel	0.49
Tone	0.96

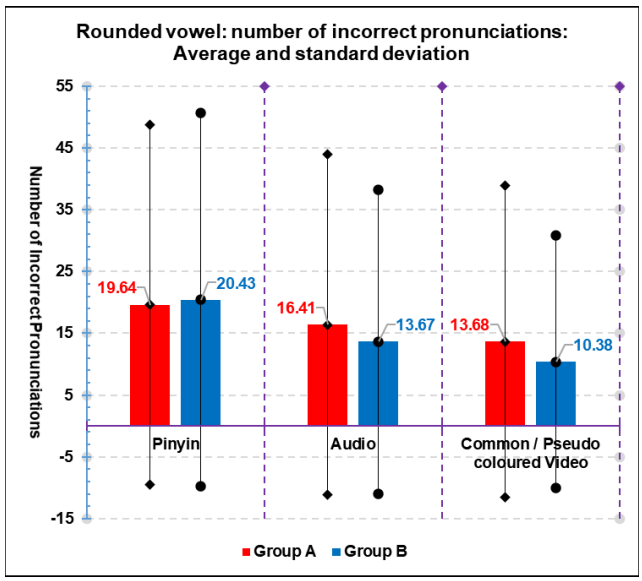

Fig. 9 Overall performance concerning the number of incorrect pronunciations of rounded vowels

sults on rounded vowels, and next, we briefly show the results regarding unrounded vowels, consonants and tones for comparison.

Figure 9 shows the overall result for rounded vowels. The graph shows the number of average and standard deviations of incorrect pronunciations. The left and middle columns show the results of the cases in which pinyin symbol material (a) and audio material (b) were presented to participants, respectively. The third column shows the results of presenting a conventional video (c) and a pseudo-coloured video (d) to Group A and Group B, respectively. From the average values, we can see that pronunciations made good improvements in the case of both the conventional video and the pseudocoloured video. They show the superiority of video-based materials over the pinyin symbol and audio-based materials.

Next, we compare the effects of conventional videos and pseudocoloured videos. A direct comparison is not possible because the performances of the participants are slightly different between the two groups.

For this purpose, we use regression analysis as shown in Fig. 10. The graph shows the estimation of how many potentially incorrect pronunciations using pinyin symbol materials also occur with video-based materials. The horizontal axis represents the number of incorrect pronunciations with pinyin symbol materials, and the vertical axis represents the number of incorrect pronunciations with video-based materials. Each dot represents the result of one participant. The

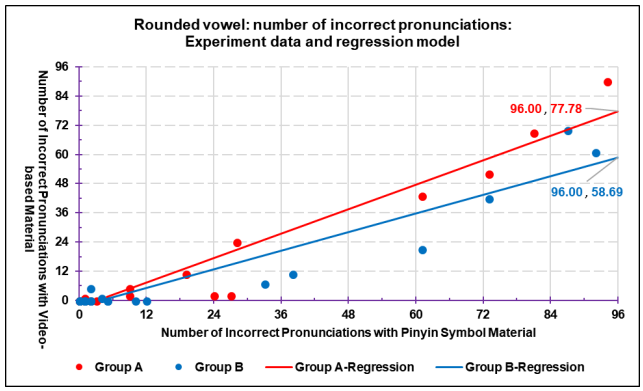

Fig. 10 Experiment data and regression models regarding rounded vowels

regression indicates how many incorrect pronunciations are expected to occur by using each video-based material. In other words, a smaller inclination means a larger amount of potential improvements from the baseline method.

The red line ($Y = 0.84 \times X - 2.80$) and the blue line ($Y = 0.64 \times X - 2.68$) in Fig. 10 are the linear regressions for Group A and Group B, respectively, where Y indicates the number of incorrect pronunciation for rounded vowels through the video-based learning materials-material type (c) or (d); X represents the number of incorrect pronunciation for rounded vowels through the pinyin learning material-material type (a), that is the baseline method in this comparison. The inclination for Group A is greater than it for Group B, ($p < 0.01$ for both cases). The R-squared values of Group A and Group B are 0.93 and 0.89, respectively, both of which show the reliability of the regression.

Similar effects were observed when we consider audio learning material-material type (b) as the baseline method. Linear regressions are $Y = 0.91 \times X - 1.18$ and $Y = 0.82 \times X - 0.82$ for Group A and Group B, respectively. Group B has smaller inclination value than Group A ($p < 0.01$ for both cases) too. The R-squared values are 0.97 and 0.98 for Group A and Group B, respectively. Both the results demonstrated that more improvement can be expected through pseudocoloured video material than common video material.

A t-test for rounded vowels using conventional video materials and pseudocoloured video materials was also conducted. In this calculation, we excluded the data of participants who made correct pronunciations for almost all samples with pinyin symbol materials. Those participants obtained almost perfect results for both video-based methods, and improvements are not different between them, i.e. their values are either 0 or very small. To avoid this effect, we gathered the data of participants who delivered incorrect pronunciation for more than five samples, i.e. who have enough room for improvement. For those data, the t-test result shows a significant difference between the two groups (one-tail P-value is 0.0083). Based on these facts, we conclude that pseudocoloured video materials have superiority over conventional video materials.

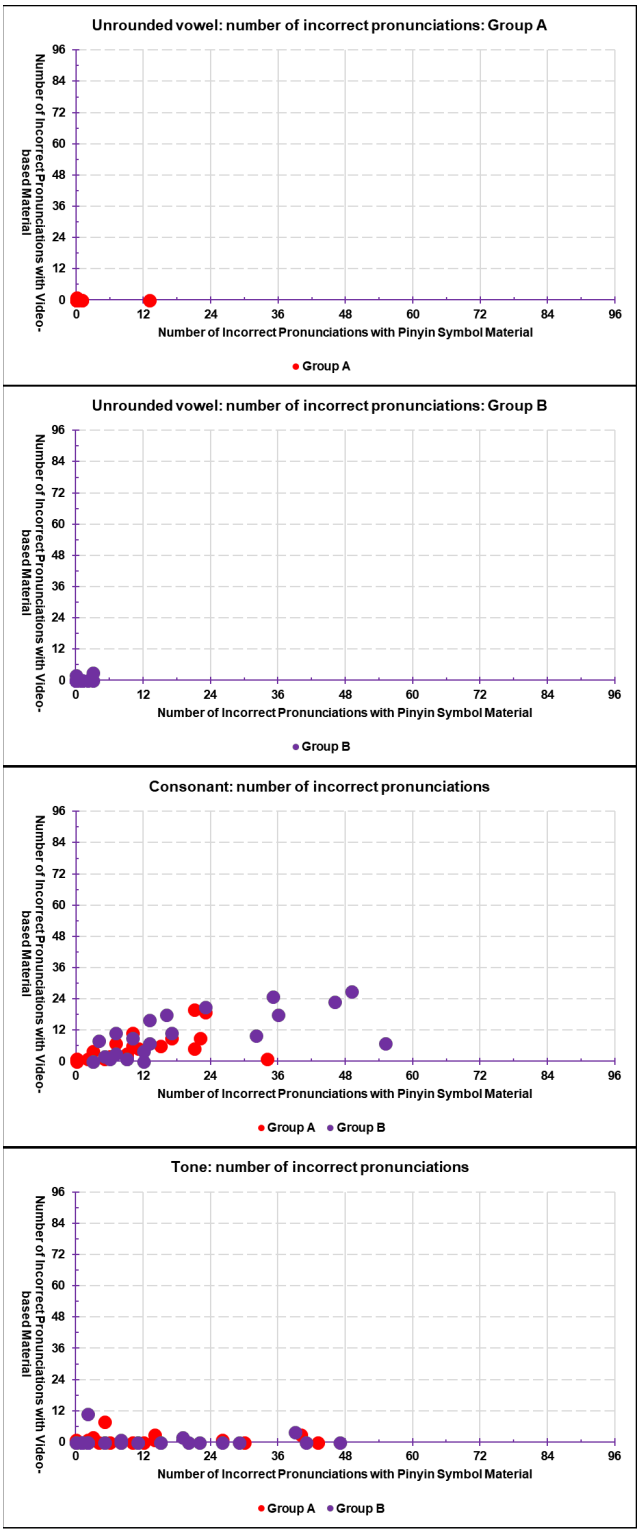

Fig. 11 Experiment data regarding unrounded vowels: Group A, unrounded vowels: Group B, consonants and tones

Figure 11 shows the experimental data distribution for unrounded vowels, consonants and tones. Since no reliable regression models were obtained, only experimental data were drawn. In contrast to rounded vowels, the results do not show any superiority of pseudocoloured video materials. For unrounded vowels, Group A and Group B were drawn separately for better viewing. All participants demonstrated good performance, and we cannot see any differences. For consonants and tones, results are more scattered; however, there are no clear performance differences between the two video-based methods. It is reasonable because a pseudocoloured video does not provide better information than a conventional video in showing consonant and tone. This implies that the differences between the videos only work for rounded vowel improvement as we had designed. The result also implies that pseudocolouring does not have a negative effect on learning consonant and tone pronunciation.

6. Conclusion

In this research, we focused on learning support for pronunciation, particularly lip protrusion. To achieve this, we proposed the pseudocolouring of face images through sensing with an RGB-D camera. This visualisation provides learners with an intuitive sense and comprehensible information regarding lip protrusion. This method can be used in preparation of teaching materials and can also be used to check learners' pronunciations. We conducted experiments to verify the former usage. The results of the experiment indicate that our proposed method provides better performance than does the conventional video method in the reduction of incorrect vowel pronunciations. Moreover, the results show that beginners who often make mistakes demonstrate significant improvements with our method. It suggests that the method works for beginners from the early days of learning.

For future studies, we need to verify how the pseudocolouring of learners' videos helps their learning. This function will provide a new kind of visual feedback to learners. Moreover, the data can be stored in an e-portfolio and can be used for formative assessment of both teachers and students. For this purpose, we need the automatic adjustment of pseudocolouring parameters for each user. In particular, in our experiments, we determined the colouring parameters for a specific person; however, this adjustment needs to be automated for each learner. As another extension of our work, the automatic evaluation of pronunciation is a good target, whereby feedback such as pronunciation instruction can be provided. This could be a good idea for the future design of language learning environments.

References

[1] R. Blake, "Technology and the four skills," Language Learning and Technology, vol.20, no.2, pp.129–142, 2016.
[2] P. Hubbard, Computer Assisted Language Learning: Critical Concepts in Linguistics, Routledge, 2009.
[3] M. Levy and G. Stockwell, CALL dimensions: Options and issues in computer-assisted language learning, Lawrence Erlbaum Associates, 2006.
[4] M. Celce-Murcia, "Teaching english as a second or foreign language 3rd edition," 2001.
[5] L. Ménard, C. Toupin, S.R. Baum, S. Drouin, J. Aubin, and M. Tiede, "Acoustic and articulatory analysis of french vowels produced

by congenitally blind adults and sighted adults," The Journal of the Acoustical Society of America, vol.134, no.4, pp.2975–2987, 2013.
[6] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," Nature, vol.264, no.5588, pp.746–748, 1976.
[7] S.-H. Hew and M. Ohki, "Effect of animated graphic annotations and immediate visual feedback in aiding japanese pronunciation learning: A comparative study," CALICO Journal, vol.21, no.2, pp.397–419, 2004.
[8] Y. Tsubota, M. Dantsuji, and T. Kawahara, "An english pronunciation learning system for japanese students based on diagnosis of critical pronunciation errors," ReCALL, vol.16, no.1, pp.173–188, 2004.
[9] R. Duan, J. Zhang, W. Cao, and Y. Xie, "A preliminary study on asr-based detection of chinese mispronunciation by japanese learners," Fifteenth Annual Conference of the International Speech Communication Association, 2014.
[10] L. Yang, L. Zhang, H. Dong, A. Alelaiwi, and A. El Saddik, "Evaluating and improving the depth accuracy of kinect for windows v2," IEEE Sensors J., vol.15, no.8, pp.4275–4285, 2015.
[11] E. Lachat, H. Macher, M.-A. Mittet, T. Landes, and P. Grussenmeyer, "First experiences with kinect v2 sensor for close range 3D modelling," The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol.40, no.5, pp.93–100, 2015.
[12] O. Wasenmüller and D. Stricker, "Comparison of kinect v1 and v2 depth images in terms of accuracy and precision," Asian Conference on Computer Vision, pp.34–45, Springer, 2016.
[13] R.C. Gonzalez and R.E. Woods, Digital Image Processing second edition, Prentice Hall, 2002.
[14] D.T. Campbell and J.C. Stanley, "Experimental and quasi-experimental designs for research," Handbook of research on teaching, pp.171–246, 1963.
[15] D.B. Rubin, "Inference and missing data," Biometrika, vol.63, no.3, pp.581–592, 1976.
[16] R. Little and D. Rubin, Statistical Analysis with Missing Data, Wiley Series in Probability and Statistics, Wiley, 2014.
[17] K.A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," Tutorials in quantitative methods for psychology, vol.8, no.1, pp.23–34, 2012.
[18] R.J. Light, "Measures of response agreement for qualitative data: Some generalizations and alternatives.," Psychological bulletin, vol.76, no.5, pp.365–377, 1971.
[19] J.R. Landis and G.G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol.33, no.1, pp.159–174, 1977.

Appendix:

Calculation for HSI values:

Converting RGB to HSI [13], we get the following:

$$H = \begin{cases} \theta & B \le G \\ 360 - \theta & B > G \end{cases}$$

With

$$\theta = \cos^{-1}\left\{\frac{1/2\left[(R-G)+(R-B)\right]}{\left[(R-G)^2+(R-B)(G-B)\right]^{1/2}}\right\}$$

$$S = 1 - \frac{3}{R+G+B}\left[min\left(R,G,B\right)\right]$$

$$I = \frac{1}{3}\left(R+G+B\right)$$

Converting HSI to RGB [13], we get the following:

Red-Green sector $0 \le H < 2\pi/3$:

$$\begin{aligned} B &= I\left(1-S\right) \\ R &= I\left[1+\frac{S\cos H}{\cos\left(\pi/3-H\right)}\right] \\ G &= 3I-\left(R+B\right) \end{aligned} \tag{A·1}$$

Green-Blue sector $2\pi/3 \le H < 4\pi/3$:

$$\begin{aligned} R &= I\left(1-S\right) \\ G &= I\left[1+\frac{S\cos\left(H-2\pi/3\right)}{\cos\left[\pi/3-\left(H-2\pi/3\right)\right]}\right] \\ B &= 3I-\left(R+G\right) \end{aligned} \tag{A·2}$$

Blue-Red sector $4\pi/3 \le H \le 2\pi$:

$$\begin{aligned} G &= I\left(1-S\right) \\ B &= I\left[1+\frac{S\cos\left(H-4\pi/3\right)}{\cos\left[\pi/3-\left(H-4\pi/3\right)\right]}\right] \\ R &= 3I-\left(G+B\right) \end{aligned} \tag{A·3}$$

Hue determination according to distance segmentations gives the following:

Lip subarea:

$\Delta d \le 20$ (upper lip subarea):

$$\tilde{H} = \begin{cases} 0 & \Delta d < 0 \\ \Delta d & 0 \le \Delta d \le 20 \end{cases}$$

$\Delta d \le 27$ (lower lip subarea):

$$\tilde{H} = \begin{cases} 0 & \Delta d < 0 \\ \Delta d & 0 \le \Delta d \le 27 \end{cases}$$

$20 < \Delta d \le 60$ (upper lip subarea):

$$\tilde{H} = (\Delta d - 20) + 120$$

$27 < \Delta d \le 60$ (lower lip subarea):

$$\tilde{H} = (\Delta d - 27) + 120$$

Non-lip subarea:

$\Delta d \le 60$:

$$\tilde{H} = \begin{cases} 120 & \Delta d < 0 \\ \Delta d + 120 & 0 \le \Delta d \le 60 \end{cases}$$

Lip and non-lip subarea:

$60 < \Delta d \le 100$:

$$\tilde{H} = (\Delta d - 60) + 240$$

$100 < \Delta d$: Original value that is converted from RGB.

The maximum allowed saturation is calculated on the basis of the variation of formula (A·1), (A·2) and (A·3). For example, when the hue is around red, the maximum allowed saturation calculation is as follows:

$$\tilde{S}_{max-red} = \frac{\left(\tilde{R}_{max}/\tilde{I}-1\right)\cos\left(\pi/3-\tilde{H}\right)}{\cos\tilde{H}}$$

where $\tilde{R}_{max}$ is set to the maximum allowed value, i.e. 1.0.

Siyang Yu received the B.S. and M.S degree in Educational Technology from Beihua University and Northeast Normal University, China, in 2006 and 2010, respectively. He is currently working toward the Ph.D. degree in Graduate school of Engineering, Kyoto University. His current research interests include e-learning supportive system and intelligent CAI.

Kazuaki Kondo received his M.E. and Ph.D. degrees from Osaka University in Japan. He became a research associate at Osaka University in 2007, an assistant professor at Kyoto university in 2009, and a lecturer in 2015. He was awarded the Kusumoto award in 2002. His research interests are computer vision and intelligent support on human communications. He is a member of IEICE.

Yuichi Nakamura received B.E, M.E, and Ph.D degrees in electrical engineering from Kyoto University, in 1985, 1987, and 1992, respectively. From 1990 to 1993, he worked as an instructor at the Department of Electrical Engineering of Kyoto University. From 1993 to 2004, he worked for Institute of Information Sciences and Electronics of University of Tsukuba, Institute of Engineering Mechanics and Systems of University of Tsukuba, as an assistant professor and an associate professor, respectively. Since 2004, he has been a professor of Academic Center of Computing and Media Studies, Kyoto University. His research interests are on computer vision, multimedia, human-computer and human-human interaction including distance communication, and multimedia contents production.

Takayuki Nakajima received the B.S. degree in Human Institute and M.S. degrees in Human Environmental Studies from Kyoto University in 2012 and 2014, respectively. Since 2014, he has managed in self access center for language learning under Professor Dantsuji.

Hiroaki Nanjo received the B.E. degree in 1999, the M.E. degree in 2001, and the Ph.D. degree in 2004 from Kyoto University. After receiving his Ph.D degree, he was a Research Associate and an Assistant Professor at Ryukoku University. From 2015/08, he is an Associate Professor at Kyoto University.

Masatake Dantsuji received the B.S. and M.S. degrees in Letters from Kyoto University in 1979 and 1981, respectively. During 1990–1997, he stayed in Kansai University as associate professor. From 1997, he is a professor of Kyoto University.