

Profile Lip Reading for Vowel and Word Recognition

Takeshi Saitoh

Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka, 820-8502, Japan
saitoh@ces.kyutech.ac.jp

Ryosuke Konishi

Tottori University
4-101 Koyama-minami, Tottori, 680-8552, Japan

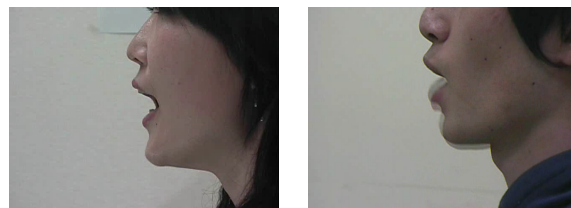
Abstract

This paper focuses on the profile view, which is the second most typical angle after the frontal face, and proposes a profile view lip reading method. We applied the normalized cost method to detect profile contour. Five feature points, the tip of the nose, upper lip, lip corner, lower lip, and chin, were detected from the contour, and eight features obtained from the five feature points were defined. We gathered two types of utterance scenes, five Japanese vowels and 20 Japanese words. We selected 20 combinations based on the eight features and carried out recognition experiments. Recognition rates of 99% for vowel recognition and 86% for word recognition were obtained with five features: two lip heights, two protrusion lengths, and one lip angle.

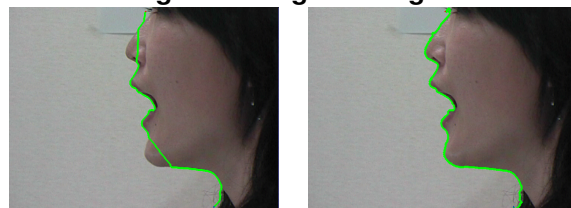
1. Introduction

Much of the research on lip reading uses frontal facial images [4, 6, 7, 9, 10]. A frontal facial image contains much useful information. However, when using a cellular phone, the difficulty of taking a good picture becomes a problem. It can be imagined that in the video-phone mode of a cellular phone, the user can adjust the position of the camera manually based on the mouth's reflection to take a picture of the frontal face. The microphone has a wide acquisition range for acquiring auditory information and the installation location is not limited. On the other hand, the location for taking a frontal facial image for traditional lip reading is limited. It is necessary to assume the situations in which a frontal facial image cannot be acquired for lip reading.

Thus, this paper focuses on the profile image as shown in Fig. 1, which is the second most typical angle after the frontal face. This image can be acquired with a camera embedded near the microphone of a cellular phone. Some studies have examined profile view



(a) (b)
Figure 1. Original images.



(a) applied result of IS (b) applied result of NC

Figure 2. Facial contour extracted results using IS and NC.

lip reading [1, 2, 3]. Most of them target words. In this paper, recognition experiments using single Japanese sounds and words are carried out, and the effectiveness of the proposed method is evaluated.

The process flow of the proposed method is as follows: It is important to extract facial contours automatically. In this method, first, two points of the upper and lower lips are detected automatically. Next, the normalized cost (NC) method is applied, which extracts object contours between two seed points, to extract the facial contour. Then, five feature points are detected. Moreover, eight features based on the five feature points are calculated. Finally, these features are fed to HMM.

2. Profile lip reading method

2.1. IS method and NC method

For extracting an object boundary, there is a well-known method by Mortensen and Barrett [5] called "in-

telligent scissors” (IS). See [5] for the details of IS. IS is a manual method for drawing a boundary based on a number of manually selected points on a visually identified boundary, providing a route that minimizes the sum of local costs. It is fast because dynamic programming is employed. However, it requires many manually selected points, particularly for an object with a complicated boundary. Figure 2(a) shows the result of applying IS when two seed points were given in Fig. 1(a). In this case, the forehead and neck on the facial contour were given as seed points. IS tends to prefer a route with a shorter path rather than a longer path because the total cost can be minimized with a shorter path.

Saitoh and Kaneko proposed a new route search method similar to IS called the NC method [8]. This method searches for a boundary that minimizes the normalized cost, i.e., sum of local costs divided by route length. The idea of NC is to consider only a partial average cost from all the past pixels within a distance of m from the coordinate of the current pixel. The route with the lowest associated cost is selected from many alternative routes between two end points as in IS. See [8] for the details of NC. The advantage of NC is the ability to extract a complex boundary using a small number of seed points. Figure 2(b) shows the result of applying NC using the two same seed points, as in Fig. 2(a). NC searches for the lowest normalized cost and can obtain a complex route, unlike IS.

2.2. Facial contour extraction

Two seeds are required to apply NC. Here, the upper and lower sides of the profile image are parts of the facial contour. For these positions, the forehead and neck, which move little during utterance, are observed. The edge of the facial contour can be detected stably by an edge detection method. Then, we apply the Sobel filter to the upper and lower sides, and the position having high edge value is detected as a seed of NC.

2.3. Facial points detection

Kumar et al. detected four facial points: the tip of the nose, the upper lip, the lower lip, and the chin from facial contours. They hypothesized these points to be local maxima in the profile contour. However, the chin, for example, is not a local maximum. Then, they applied a distance transformation that maps the original profile contour to a new contour called T -transformed. This method uses the maximum values of the x and y axes, and it is necessary to limit the range of the face that is reflected in the image. Thus, we do not applied T -transformed, but calculate the curvature of the facial contour and detect five facial points, the tip of the nose P_1 , the upper lip P_2 , the lip corner P_3 , the lower lip

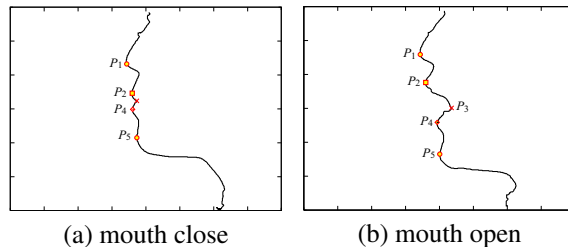


Figure 3. Facial feature points.

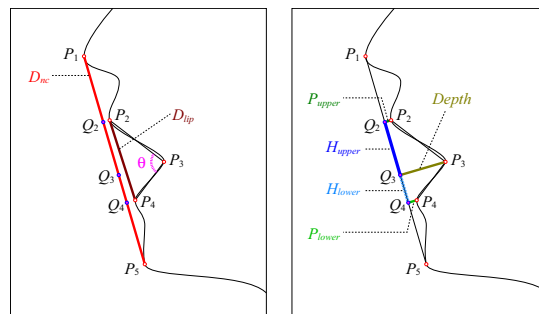


Figure 4. Profile features.

P_4 , and the chin P_5 . The five features that are detected when the mouth closes and opens are shown in Fig. 3.

2.4. Profile features

This paper defines the following eight features. The relationship between facial points and features is shown in Fig. 4. (1) We define the Euclidean distance between the points of the upper lip P_2 and the lower lip P_4 as the feature D_{lip} . (2) We define the Euclidean distance between the points of the nose P_1 and the chin P_5 as the feature D_{nc} . (3) The intersection point of two straight lines is defined for the perpendicular from upper lip P_2 to the straight line between the nose and the chin as Q_2 . Then, we define the Euclidean distance between points P_2 and Q_2 as the protrusion length P_{upper} . (4) As for the protrusion length of the upper lip, the intersection point of two straight lines is defined for the perpendicular from lower lip P_4 to the straight line between the nose and the chin as Q_4 . Then, we define the Euclidean distance between points P_4 and Q_4 as the protrusion length P_{lower} . (5) The intersection point of two straight lines is defined for the perpendicular from P_3 to the straight line between the nose and the chin as Q_3 . Then, we define the Euclidean distance between points P_3 and Q_3 as the lip $Depth$. (6) We define the Euclidean distance between points Q_2 and Q_3 as the height of the upper lip H_{upper} . (7) We define the Euclidean distance between points Q_3 and Q_4 as the height of the lower lip H_{lower} . (8) This feature quantitatively expresses the amount by which the lips are open and θ is defined as an angle of the lip.

P_{upper} , P_{lower} , $Depth$, H_{upper} , H_{lower} , and θ were

used by the method in [2].

2.5. Recognition method

HMM is a well-known method for recognizing time series data, and is applied by many of the studies [1, 2, 3] concerning lip reading. This research also applies HMM for the recognition process.

3. Experiment

A recognition experiment was carried out with 20 combinations of utterance based on eight features calculated from profile views of utterance and the effective features were determined. Here, we used an HMM toolkit (HTK).

3.1. Vowel recognition

In the vowel recognition experiment, we collected 20 utterance scenes of each vowel from five persons (A, B, C, D, and E) out of which three persons (A, B, and C) were men. Thus, we took 100 utterance scenes per person. The five target Japanese vowels were /a/, /i/, /u/, /e/, and /o/. The image size was 640×480 pixels and the frame rate was 30 frames per second. A person sat on a chair with a straight posture and a fixed digital video camera was focused on his face. He/She closed his/her lips before and after each utterance. All scenes were taken in the same environment. For each person, we divided the 20 samples into two groups of 19 samples for training and 1 for recognizing, i.e., we applied the leave-one-out method. Our recognition method is based on the training data for each utterance scene. Thus, speaker-dependent speech recognition is targeted here.

The resulting average recognition rates are shown in Table 1. In this table, F_i , $i = 1, 2, \dots, 20$, is the combination number of the features, and n is the number of features. In $n = 1$, F7 (depth of the lip) obtained the highest recognition rate of 88.0%. In $n = 2$, F11 (two height features) obtained the highest recognition rate of 96.8%. F18 (four features of the height and protrusion of lip) obtained the overall highest recognition rate of 99.6%. F11 is the same feature combination as in Kumar’s research. It was confirmed to obtain a recognition accuracy of 95% or more with $n > 2$.

3.2. Word recognition

In the word recognition experiment, we set the 20 Japanese words shown in Table 2 as the target. These words are commonly used in telephonic conversations. We collected ten utterance scenes of each word from three persons (A, B, and F, all men). Thus, we took 200 utterance scenes per person. A and B are same person of the previous experiment. As well as other conditions such as image size, frame rate, and photography conditions were the same as in the previous experiment.

Table 1. Recognition result of vowel.

feature	n	A	B	C	D	E	ave
F1: D_{lip}	1	90	74	90	78	62	78.8
F2: D_{nc}	1	86	58	88	80	68	76.0
F3: P_{upper}	1	96	60	52	80	84	74.4
F4: P_{lower}	1	68	52	80	86	82	73.6
F5: H_{upper}	1	82	68	84	80	94	81.6
F6: H_{lower}	1	68	76	82	76	62	72.8
F7: $Depth$	1	86	82	96	88	88	88.0
F8: θ	1	84	66	64	82	94	78.0
F9: D_{lip}, D_{nc}	2	98	86	94	90	98	93.2
F10: P_{upper}, P_{lower}	2	100	82	90	86	98	91.2
F11: H_{upper}, H_{lower}	2	98	98	92	96	100	96.8
F12: $D_{lip}, Depth$	2	98	84	100	98	94	94.8
F13: $Depth, \theta$	2	96	72	98	96	90	90.4
F14: $H_{upper}, H_{lower}, \theta$	3	100	94	98	98	100	98.0
F15: $P_{upper}, P_{lower}, \theta$	3	100	98	94	96	100	97.6
F16: $D_{lip}, H_{upper}, H_{lower}$	3	100	90	96	92	100	95.6
F17: $D_{lip}, P_{upper}, P_{lower}$	3	100	96	98	100	100	98.8
F18: $H_{upper}, H_{lower}, P_{upper}, P_{lower}$	4	100	100	98	100	100	99.6
F19: $D_{lip}, P_{upper}, P_{lower}, \theta$	4	100	98	98	100	100	99.2
F20: $H_{upper}, H_{lower}, P_{upper}, P_{lower}, \theta$	5	100	98	98	100	100	99.2

The resulting average recognition rates are shown in Table 3. The accuracy of word recognition is 10-30% less than that of vowel recognition, though it differs according to the combination of the features examined. In $n = 1$, F7 obtained the highest recognition rate of 72.3%, as it did for vowel recognition. F20 (five features of height, protrusion, and angle of lip) obtained the overall highest recognition rate of 85.7%.

3.3. Human lip reading versus our method

To evaluate the recognition results, we carried out subject experiments and analyzed the recognition accuracy by comparing these results and the computational results. In subject experiments, a subject watches an utterance scene two times and identifies the content of utterance by visual observation. In this experiment, the utterance scene is chosen at random from the scenes used in 3.1 and 3.2 for the five vowels or 20 words.

The subject experiment was carried out 20 times per subject without prior training. Eighty-four students, out of which six were women, cooperated in this experiment. The resulting recognition rates for the five vowels and 20 words were 80.0% and 67.3%, respectively. These results indicate that our method is more accurate than visual observation.

3.4. Discussion

The utterance of a word is an active movement, whereas the utterance of a vowel is a monotonous movement. F20 obtained the highest recognition rate of 94.1% when the average recognition results for vowels and words were considered. Thus, we calculated two confusion matrices for F20. The confusion matrix contains information about the actual and predicted result of the recognition task. Regarding vowel recognition, /e/ and /o/ were misidentified as /a/ and /u/, respectively. However, the false recognition rate was only 2%. On the other hand, for word recognition, w15 was misrecognized as w20 17% of the time,

Table 2. Target 20 Japanese words.

symbol	word
w01	/ko/n/ni/chi/wa/
w02	/o/ha/yo/u/
w03	/o/ya/su/mi/na/sa/i/
w04	/a/ri/ga/to/u/
w05	/o/me/de/to/u/
w06	/i/ku/
w07	/ku/ru/
w08	/o/shi/e/ru/
w09	/wa/ka/ru/
w10	/a/so/bu/
w11	/ta/be/ru/
w12	/no/mu/
w13	/do/ko/
w14	/ji/ka/n/
w15	/ba/syo/
w16	/i/ma/su/ka/
w17	/i/ma/se/n/
w18	/i/i/de/su/ka/
w19	/shi/te/ku/da/sa/i/
w20	/ma/ta/a/i/ma/syo/u/

Table 3. Recognition result of word.

feature	n	A	B	F	ave
F1: D_{lip}	1	71.5	59.0	38.5	56.3
F2: D_{nc}	1	53.5	72.0	29.0	51.5
F3: P_{upper}	1	80.5	45.5	54.5	60.2
F4: P_{lower}	1	81.0	45.0	44.5	56.8
F5: H_{upper}	1	83.0	56.5	45.0	61.5
F6: H_{lower}	1	49.5	47.0	39.5	45.3
F7: $Depth$	1	91.0	64.5	61.0	72.3
F8: θ	1	72.5	45.0	57.5	58.3
F9: D_{lip}, D_{nc}	2	79.5	81.0	46.0	68.8
F10: P_{upper}, P_{lower}	2	87.5	63.5	60.5	70.5
F11: H_{upper}, H_{lower}	2	80.5	64.5	53.0	66.0
F12: $D_{lip}, Depth$	2	97.5	81.0	76.5	85.0
F13: $Depth, \theta$	2	96.0	73.5	70.5	80.0
F14: $H_{upper}, H_{lower}, \theta$	3	91.5	77.0	69.5	79.3
F15: $P_{upper}, P_{lower}, \theta$	3	91.5	72.5	72.0	78.7
F16: $D_{lip}, H_{upper}, H_{lower}$	3	81.5	67.0	44.5	64.3
F17: $D_{lip}, P_{upper}, P_{lower}$	3	93.0	81.5	67.5	80.7
F18: $H_{upper}, H_{lower}, P_{upper}, P_{lower}$	4	96.0	81.0	72.5	83.2
F19: $D_{lip}, P_{upper}, P_{lower}, \theta$	4	96.0	82.0	74.5	84.2
F20: $H_{upper}, H_{lower}, P_{upper}, P_{lower}, \theta$	5	98.0	83.0	76.0	85.7

and w17 was misrecognized as w16 13% of the time. The first two characters of w16 (/i/ma/su/ka/) and w17 (/i/ma/se/n/) are the same, hence mistaking one for the other is considered understandable. In contrast, w15 and w20 are different words. We investigated the reason in detail and found that 80% of the false recognition of w15 and w20 was done by B. The number of utterance frames of B were 66.2 and 76.7, respectively, though the number of characters in w15 and w20 were two and seven, respectively. In a word, there is no difference in the number of frames, though there is a difference in the number of characters. Moreover, the utterance of /ba/syo/ and /ma/syo/u/ involves similar lip movements. The false recognition was thought to have occurred for these reasons.

4. Conclusion

This paper has made two major contributions. The first is automatic profile contour detection based on normalized cost method. Some traditional research used

markers or a blue background to detect the profile views easily. The second is the experiments on single sounds and words. We set five Japanese vowels as the target of single sound recognition and 20 Japanese words as the target of word recognition. As a result, five features (two height values, two protrusion values, and one lip angle) yielded high recognition accuracy for both the targets. Moreover, to verify recognition accuracy, we conducted an additional subject experiment with 84 students and found that computer-based recognition accuracy is higher than that of humans.

Acknowledgement

This research was partially supported by MEXT Grant-in-Aid for Young Scientists (B) 21700582, JSPS Grant-in-Aid for Scientific Research (C) 21500515, and SCOPE (Strategic Information and Communications R&D Promotion Programme) of MIC.

References

- [1] K. Iwano, T. Yoshinaga, S. Tamura, and S. Furui. Audio-visual speech recognition using lip information extracted from side-face images. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007:69–73, 2007. doi:10.1155/2007/64506.
- [2] K. Kumar, T. Chen, and R. M. Stern. Profile view lip reading. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, number 4, pages 429–432, 2007.
- [3] P. J. Lucey and G. Potamianos. Lipreading using profile versus frontal views. In *Proc. of IEEE 8th Workshop on Multimedia Signal Processing*, pages 24–28, 2006.
- [4] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey. Extraction of visual features for lipreading. *IEEE Trans. Pattern Anal. & Mach. Intell.*, 24(2):198–213, 2002.
- [5] E. N. Mortensen and W. A. Barrett. Intelligent scissors for image composition. In *SIGGRAPH 1995*, pages 191–198, 1995.
- [6] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audio-visual speech. In *Proc. of IEEE*, volume 91, pages 1306–1326, 2003.
- [7] T. Saitoh, M. Hisagi, and R. Konishi. Analysis of features for efficient japanese vowel recognition. *IEICE Trans. Inf. & Syst.*, E90-D(11):1889–1891, 2007.
- [8] T. Saitoh and T. Kaneko. Route search method by normalized cost. *Systems and Computers in Japan*, 36(14):11–20, 2005.
- [9] T. Saitoh and R. Konishi. Japanese 45 single sounds recognition using intraoral shape. *IEICE Trans. Inf. & Syst.*, E91-D(11):2735–2738, 2008.
- [10] T. Saitoh, K. Morishita, and R. Konishi. Analysis of efficient lip reading method for various languages. In *Proc. of International Conference on Pattern Recognition (ICPR2008)*, 2008.