MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

Japanese Phone Recognition using Lip Image Information

Takeshi Saitoh, Mitsugu Hisagi and Ryosuke Konishi
Department of Electrical and Electronic Engineering, Tottori University
4-101 Koyama-minami, Tottori-shi, Tottori 680-8552 Japan
E-mail: saitoh@ele.tottori-u.ac.jp

Abstract

This paper describes Japanese phone recognition for lip reading based on a novel feature called trajectory feature to obtain high recognition rate. Trajectory feature is a time change of two mouth shape features expressed as a two-dimensional trajectory of the lip motion. The most similar trajectory in a database which is compared with the target trajectory by DP matching is selected as a result phone. Traditional researches concerned phone recognition experimented with only Japanese five vowels. In this paper, experiments were conducted for Japanese 45 phones with five person, and classified into five vowels and ten consonants, and recognition rate of 94.1% and 28.9%, respectively, were obtained.

1 Introduction

When perceiving speech, humans not only use the auditory information but also other information, such as lip motion, eye gaze, hand gestures, etc. Since the recognition rate using auditory information decreases by the surrounding noise. However other information does not have an influence with the noise. The importance of this information becomes significant in noisy environments.

Recently, speech, especially word recognition using visual information called lip reading, has attracted significant interest. In lip reading, there are roughly three target, phone recognition, word recognition, and continuous speech recognition. However, lip reading is a more difficult task than speech recognition with auditory information. A lot of researches concerned with lip reading are targeted in word recognition [1, 2, 3, 4] or with auditory information [5, 6]. Since phone scene contains few phonemes, vowel recognition rate is lower than that of the word. Some researchers have experimented in vowel recognition [7, 8, 9, 10, 11]. They tried only the recognition of about Japanese five vowels. This research conducts Japanese 45 phones, and tries to classify these phones into five vowel classes and ten consonant classes. This paper also proposes a novel feature to obtain high recognition rate for lip reading.

Lyons et al. proposed a text entry method, which uses coordinated motor action of the key input by hand and the mouth shape [12]. Two shape parameters (the area and the aspect ratio) of the open mouth cavity are gauged. Matsuoka et al. showed a way to distinguish the vowels, with a front and a side view [7]. Nakamura et al. applied an Active Contour Model to extract the lip shape, and the extracted contour points are fed to a neural network to estimate Japanese five vowels [10].

Our lip reading method followed is: First, region of interest of lip is detected during the input image sequences. Moreover, the mouth cavity region is detected with simple thresholding method. The lip size is then normalized to prevent from influencing the features. Here, a user closes a lip before and after the utterance, and a speaking period is detected. Next, a time change of two lip shape features (area and aspect ratio) is expressed as a two-dimensional trajectory of the lip motion. The most similar trajectory in a database which is compared with the target trajectory by DP matching is selected as a result phone.

2 Mouth cavity detection

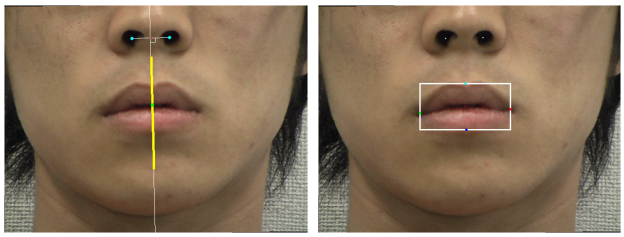
2.1 ROI detection

There are two methods of lip reading, the image-based method [2, 6], and the model-based method [1, 3, 4]. In the image-based method, the typical feature is computed based on eigen space projected by an image around the detected lip region. It does not need to construct a special model, and information of the tooth and the tongue can be included. But a large amount of data is required. There is a large influence of the lighting conditions and the location of lip. Oppositely, in the model-based method, a contour or some target points of the lip are first required with a minimum amount of parameters. The model constructed by this method presents less influence from the lighting conditions and the lip location. It has a problem that it is difficult to construct the model, but the processing speed is better. In this paper, we first detect Region Of Interest (ROI) of lip. After ROI detection, we employ the thresholding method to obtain a mouth cavity region.

To detect a ROI, two nostrils which are darker region than surrounding skin color are detected. A mouth exists in the bottom side direction of the vertical bisector between the nostrils as shown in figure 1(a). Using this direction, we compute edge profile and detect a maximum edge point as a connected point with upper and lower lip. Next, two lip terminal points (the left and the right points) are detected using circular separability filter [13], and the ROI is set as shown in figure 1(b).

2.2 Mouth cavity detection

In the ROI, we detect a mouth cavity region with simple thresholding method. To eliminate from lighting environment, the original image is converted from RGB(red-green-blue) color into HLS(hue-lightness-saturation) color space. When L is less than 0.2, the pixel is set to the mouth cavity, otherwise the



(a) (b)

Figure 1: Mouth cavity detection.

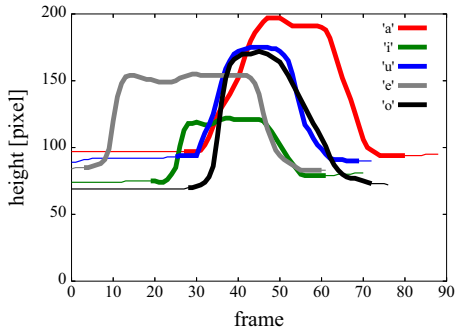


Figure 2: Results of speaking period detection.

pixel is set the lip region. The resulting image is shown in upper side of figure 3. In these images, the white box is ROI, and yellow region is the detected mouth cavity region.

2.3 Size normalization

The lip size varies with the distance between the camera and the user. Therefore, the lip size is normalized. Because a lip is closed before and after an utterance, a ratio ($R = W/W^*$) with the standard size W^* found by calculating the average width W of the lip in the first several frames is used. The normalization is applied to the target frame by using this ratio.

2.4 Speaking period

In this research, a user closes his lip before and after utterance. Then, we detect a speaking period with the height of mouth cavity region. At first, the average $\bar{h}$ of the height of several frames is calculated. In this research, we used ten frames for calculating $\bar{h}$. Then, the frame that height is bigger than a threshold ($t_h = \bar{h} + 10$) is set an utterance starting frame. An utterance finishing frame is set the frame that the difference from the average is less than t_h continuously with ten frames. Some results are shown in figure 2. In this graph, thick line is detected speaking period. The gray line ('e') is short and detected result is hard to understand. But others are clear. Particularly, the frame before utterance is ridded correctly.

3 Trajectory Feature

Because of the area S and the aspect ratio $A = H/W$ is recognized high accuracy by Lyons' experiment [12], where H is height and W is width of the mouth cavity region, we use these two features. Though Lyons et al. calculated aspect ratio with only the mouth cavity region, we defined two aspect ratio A_c and A_r, where A_c is based on mouth cavity region, A_r is based on the ROI obtained in **2.1**. Furthermore, a time change of two features is expressed as a two-dimensional trajectory of the lip motion of the target phone. The trajectory is generated by plotting points in two-dimensional space shown in lower side of figure 3.

The lower side of figure 3 shows five trajectory features of Japanese five vowels ('a', 'i', 'u', 'e', and 'o'). The horizontal axis is area S, and the vertical axis is aspect ratio A_c. Plotted red points in trajectory feature are the position of features of each frame. In this figure, small area is plotted left side and large area is plotted right side, small aspect ratio is plotted upper side and large aspect ratio is plotted lower side. The number of frames is from 35 to 91 per one phone scene which we recorded. Since there are few frames, the mis-recognition may occur when the trajectory is expressed only by the number of frames. When the trajectory is expressed with a polyline, there are many rapid changes in trajectory and it may cause mis-recognition. Hence, we express the trajectory by B-Spline curve [14]. The progress of the time is shown by the curve color. As for the color of curve, the starting color is red. Then, it changes with yellow, green, light blue, and turns blue at end.

Here, if it is the same user, the mouth shape before the utterance is almost the same. But, it does not plot the same position because it is not the same shape due to the movement such as breathing or the accuracy of region detection process, even if a lip is closed before an utterance. Thus, the starting point of the trajectory is set up in the same position to prevent a shift. In this paper, we normalize the size of trajectory feature with average value and standard deviation, then the starting point is set $(S, A) = (100, 100)$, where the size of trajectory feature is within 512×512 pixels and the lower side of figure 3 is a part of trajectory feature image.

The mouth images of figure 3 show the mouth wide open frame. These frames are mostly located in the farthest place from the starting point. In other words, the place depends on the vowel.

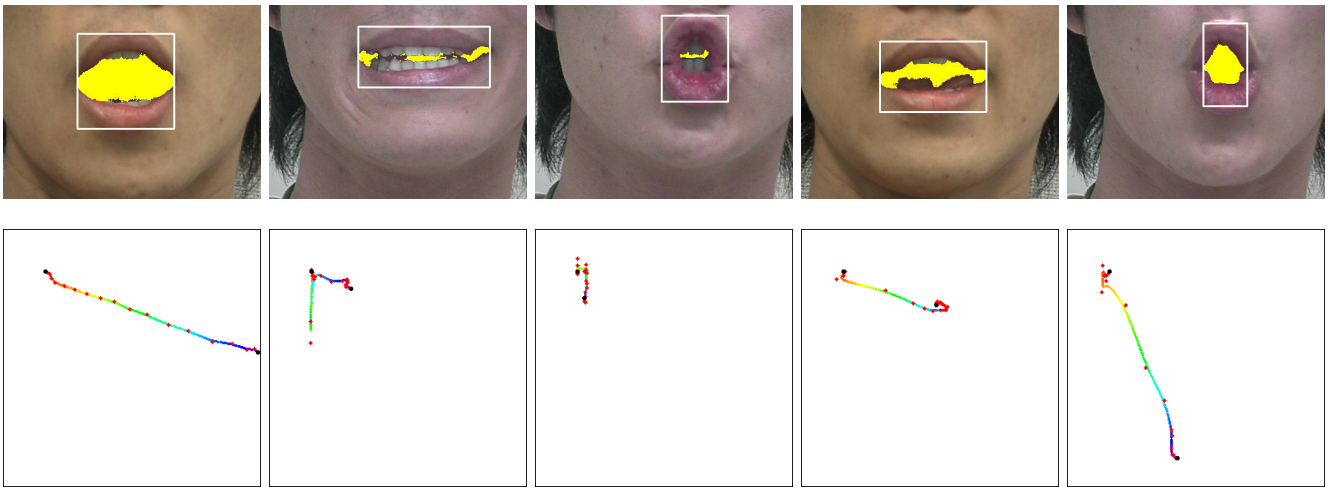
4 Recognition

DP matching is well-known method and is nonlinear matching by time warping. In this research, Two-dimensional DP matching is applied to calculate a distance $D(X, R_n)$, where $X = x_1, x_2, \cdots, x_I$ is a target unknown phone trajectory, $R_n = r_1, r_2, \cdots, r_J$ is a database trajectory for the comparison. The phone $\hat{n}$ which satisfies an eqution $\hat{n} = \arg\min_n D(X, R_n)$ is chosen. Here, x_j and r_i are two-dimensional trajectory vectors, where $x_j = (s_j^x, a_j^x)^T$ and $r_i = (s_i^r, a_i^r)^T$ consisted of an area S and an aspect ratio A. The initial value of DP matching is set as follows.

$$\begin{cases} g(i,0) = 0 & (i = 0, 1, \cdots, I) \\ g(0,j) = \infty & (j = 1, 2, \cdots, J) \end{cases}$$

An accumulation distance $g(i,j)$ in each lattice point (i,j) is calculated as following equations.

$$g(i,j) = \min \begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i,j-1) + d(i,j) \end{cases}$$



(a) 'a' (b) 'i' (c) 'u' (d) 'e' (e) 'o'

Figure 3: Mouth wide open images and trajectory features.

Table 1: 45 Japanese phones.

	vowel				
consonant	あ a	い i	う u	え e	お o
	か ka	き ki	く ku	け ke	こ ko
	さ sa	し shi	す su	せ se	そ so
	た ta	ち chi	つ tsu	て te	と to
	な na	に ni	ぬ nu	ね ne	の no
	は ha	ひ hi	ふ hu	へ he	ほ ho
	ま ma	み mi	む mu	め me	も mo
	や ya		ゆ yu		よ yo
	ら ra	り ri	る ru	れ re	ろ ro
	わ wa				を wo

Where, a local distance $d(i,j)$ is an Euclidean distance. The distance between two trajectories (R_n and X) is calculated by $D(R_n, X) = g(I,J)/(I+J)$.

5 Experiment

The basic Japanese hiragana syllabary is shown table 1. Japanese syllable structure is fairly simple in the most syllables take the form CV, where C means consonant, V means vowel, and there are only five vowels ('a', 'i', 'u', 'e', and 'o'). For experiments, we collected five samples of each phone from five persons (A, B, C, D, and E), for a total of 1125 samples. Here, B is a woman, and residual persons are men. We took image sequence of the person's clear utterance with a digital video camera. The image size is 640 × 480 pixels. The image sampling rate of inputs was 30 frames per second. The mouth cavity detection was carried out for all image sequences.

5.1 Classify into five vowels

To evaluate the performance of trajectory feature, we compared trajectory feature by DP matching with k Nearest Neighbor (k-NN) method which input is two shape features of mouth wide open frame [11]. The first experiment is to classify 45 phones into five vowels. Here, for each subject, train on four samples of each phone and classify the fifth. Thus for each person the training sets contained 180 samples (four samples for each phone) and the non-overlapping test sets contained 45 samples (the residual sample of all phones). The resulting average recognition rates are shown in table 2. The resulting recognition rate was 92.5% and 90.3% with method of [11]. And that of proposed method was higher rate, 94.1% and 93.9%. Table 3 gives a confusion matrix with (S, A_c) for five vowels. Two vowels, 'i' and 'e', have low recognition rate. To investigate this reason, figure 4 shows two distributions of 225 phones of subject D with aspect ratio and area when mouth is wide open. Figure 4(a) is with area S and aspect ratio of mouth cavity A_c, figure 4(b) is with area S and aspect ratio of lip ROI A_r. Though these distributions are almost divided clearly, 'i' and 'e' are gathered near location. However, hence, we confirmed that trajectory feature is effective in the vowel recognition.

Table 2: Classification results into five vowels.

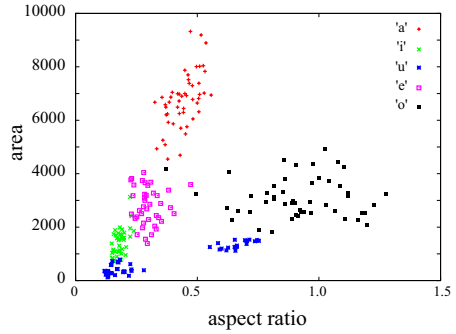
method	subject [%] A	B	C	D	E	average [%]
S, A_c + k-NN	93.3	91.1	90.2	94.2	93.8	92.5
S, A_r + k-NN	93.3	85.8	82.2	95.1	95.1	90.3
trajectory(S, A_c)	96.0	93.3	89.3	93.3	98.7	94.1
trajectory(S, A_r)	96.4	96.0	81.3	96.4	99.6	93.9

Table 3: Confusion matrix of five vowels.

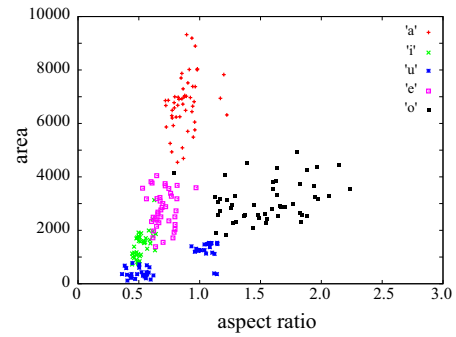
	a	i	u	e	o
a	95.8	0.0	0.0	4.0	0.2
i	0.0	91.8	2.8	5.0	0.5
u	0.0	2.9	96.9	0.0	0.2
e	4.0	6.3	0.0	88.3	1.5
o	0.8	0.4	0.6	2.0	96.2

5.2 Classify into ten consonants

Next, we carried out consonant recognition with trajectory features, that is, 45 phones were classified into ten consonants. The resulting average recognition rates are shown in table 4, and table 5 gives a confusion



(a) S and A_c.



(b) S and A_r.

Figure 4: Distributions of area and aspect ratio.

Table 4: Classification results into ten consonants.

method	subject [%] A	B	C	D	E	average [%]
trajectory(S, A_c)	27.6	22.2	23.1	24.0	47.6	28.9
trajectory(S, A_r)	28.9	26.2	23.1	27.6	47.6	30.7

Table 5: Confusion matrix of ten consonants.

	a	k	s	t	n	h	m	y	r	w
a	20.4	12.4	6.0	3.2	8.0	22.0	4.0	3.6	15.6	4.8
k	10.4	30.8	8.8	6.8	8.8	7.2	6.8	6.0	11.2	3.2
s	5.6	10.0	32.8	16.8	11.6	6.8	3.6	5.6	6.4	0.8
t	4.8	5.2	16.4	30.0	18.0	6.0	4.0	6.8	8.8	0.0
n	7.2	6.8	11.6	16.4	29.2	8.0	5.6	4.0	8.8	2.4
h	16.4	11.6	5.2	4.0	8.4	33.2	7.6	3.2	7.2	3.2
m	7.2	7.2	6.0	4.4	6.4	10.4	42.4	2.0	11.2	2.8
y	4.0	11.3	6.7	10.7	10.0	4.7	3.3	33.3	13.3	2.7
r	12.4	15.6	8.0	10.0	7.6	4.8	6.8	8.0	22.0	4.8
w	14.0	10.0	3.0	5.0	2.0	10.0	6.0	9.0	5.0	36.0

matrix with (S, A_c) for ten consonants. The recognition rates were low. These results show that the task to recognize consonant is difficult problem. Moreover, though we tried the classification for 45 phones into 45 classes, we obtained almost the same recognition result of ten consonant classifications.

6 Conclusion

This paper proposed a novel feature for obtaining high recognition rate of phone recognition for lip reading. The proposed feature is a two-dimensional trajectory that a time change of two lip shape features (area and aspect ratio). We carried out the recognition process for classifying five vowels with 45 phones, and obtained high recognition rate 94.1%. Moreover, we tried the recognition process for classifying ten consonants, and obtained recognition rate 28.9%. Consonantal classification was a low recognition rate. We consider that information inside the mouth of the tongue and the tooth is important for the consonantal recognition. Thus, the future work is consideration to imply inside information.

References

[1] K. Sugahara, T. Shinchi, M. Kishino, and R. Konishi, "Real time realization of lip reading system on the personal computer", *Journal of SICE,* (Japanese Edition), vol.36, no.12, pp. 1145–1151, 2000.

[2] J. Kim, J. Lee, and K. Shirai, "An efficient lip-reading method robust to illumination variations," *IEICE Trans. Fundamentals,* vol.E85-A, no.9, pp. 2164-2168, 2002.

[3] L. G. Silveira, J. Facon, and D. L. Borges, "Visual speech recognition: a solution from feature extraction to words classification," *SIBGRAPI 2003,* pp. 399-405, 2003.

[4] T. Saitoh and R. Konishi, "Lip reading based on sampled active contour model," *LNCS3656,* pp. 507-515, 2005.

[5] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of Visual Features for Lipreading," *IEEE Trans. on Pattern Analysis and Machine Intellignce,* vol.24, no.2, pp. 198-213, 2002.

[6] K. Murai, S. Nakamura, "Face-to-talk: audio-visual speech detection for robust speech recognition in noisy environment," *IEICE Trans. Inf. & Syst.,* vol.E86-D, no.3, pp. 505-513, 2003.

[7] K. Matsuoka, T. Furuya, and K. Kurosu, "Speech recognition by image processing of lip movements —discrimination of the vowels and its application to word recognition—", *Journal of SICE,* (Japanese Edition), vol.22, no.2, pp. 67–74, 1986.

[8] K. Uchimura, J. Michida, M. Tokou, and T. Aida, "Discrimination of Japanese vowels by image analysis", *IEICE Trans. Inf. & Syst.*, (Japanese Edition), vol.J71-D, no.12, pp. 2700–2702, Dec. 1988.

[9] J. -T. Wu, S. Tamura, H. Mitsumoto, H. Kawai, K. Kurosu, and K. Okazaki, "Neural network vowel-recognition jointly using voice features and mouth shape image", *IEICE Trans. Inf. & Syst.*, (Japanese Edition), vol.J73-D-II, no.8, pp. 1309–1314, Aug. 1990.

[10] S. Nakamura, T. Kawamura, and K. Sugahara, "Development of vowel recognition system by lip-reading method using active contour models", *Proc. of 5th Forum on Information Technology,* (Japanese Edition), pp. 353–356, 2006.

[11] M. Hisagi, T. Saitoh, and R. Konishi, "Analysis of efficient feature for Japanese vowel recognition", *Proc. of ISPACS 2006,* pp. 33–36, 2006.

[12] M. J. Lyons, C.-H. Chan, and N. Tetsutani, "Mouth Type: text entry by hand and mouth," *Proc. of Conference on Human Factors in Computing Systems,* pp. 1383-1386, 2004.

[13] M. Yuasa, O. Yamaguchi, and K. Fukui, "Precise pupil contour detection based on minimizing the energy of pattern and edge," *IEICE Trans. Inf. & Syst.,* vol.E87-D, no.1, pp. 105-112, 2004.

[14] J. D. Foley, J. F. Hughes, A. van Dam, and S. K. Feiner, "Computer Graphics Principles and Practice, " Addison-Wealey, 1993.