A Detection Method of Basic Mouth Shapes from Japanese Utterance Images

Tsuyoshi Miyazaki[1], Toyoshiro Nakashima[2], and Naohiro Ishii[3]

[1] Department of Information and Computer Sciences, Kanagawa Institute of Technology, 1030 Shimo-ogino, Atsugi, Kanagawa, Japan
[2] School of Culture-Information Studies, Sugiyama Jogakuen University, 17-3 Hoshigaoka-motomachi, Chikusa, Nagoya, Aichi, Japan
[3] Department of Information Science, Aichi Institute of Technology, 1247 Yachigusa, Yakusa, Toyota, Aichi, Japan
miyazaki@ic.kanagawa-it.ac.jp, nakasima@ci.sugiyama-u.ac.jp, ishii@aitech.ac.jp

Abstract. Some distinctive mouth shapes are formed when Japanese words and phrases are uttered. Because people who have acquired a skill of Japanese lip-reading know these characteristics, they can read lips movement. To realize the machine lip-reading, we propose a method which detects the distinctive mouth shapes from Japanese-speaking images based on their techniques. We define six mouth shapes as the distinctive mouth shapes, and the mouth shape images are used as template images. To detect the mouth shapes in utterance images, template matching is applied. Waveforms of similarity which are calculated by the template matching show some characteristic forms. Thus, we detect the mouth shapes from the waveforms. We carry out some experiments using Japanese words, and confirm effectiveness of the proposed method from the results.

Keywords: Lip-reading, Japanese phoneme, Template matching.

1 Introduction

Lip-reading is one of the important communication means for hearing-impaired people. In the lip-reading, a content of utterance is understood by movement of speaker's lips, etc. In recent years, some researches for realizing the lip-reading with information processing technologies have been pursued. This is called "machine lip-reading". It is used as a complementary technology to improve speech recognition, and the research of lip-reading for supporting communication with the hearing-impaired people is being pursued.

Generally, in the machine lip-reading, several images (frame images) of the lips region are taken using a camera or similar devices during a speaker is uttering. Digital image processing is performed on the frame images, and features about changes in mouth shape and movement of lips are calculated during an utterance. For example, a method which uses optical flow generates the features from velocity vectors[1]. A method which uses mouth shape changes generates the features from aspect ratio of

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2011, LNCS 6761, pp. 608–617, 2011.

the lips region[2,3]. In addition, a method which uses images of the lips region generates the features from the images by template matching[4].

On the other hand, we found that people who have acquired a skill of lip-reading ("lip-reading skill holder" is used hereafter) stare at the mouth shape of speakers when reading lips. In addition, some distinctive mouth shapes are formed intermittently when uttering Japanese phones[1].

Thus, as the first step for realizing the machine lip-reading by modeling the lip-reading skill holder, we proposed a method in which knowledge of the lip-reading skill holders is logically materialized and the distinctive mouth shapes are processed using computers[5]. In this proposal, we defined six mouth shapes as "Basic Mouth Shape" (BaMS), and they are /a/, /i/, /u/, /e/, /o/ and closed mouth. In addition, there are some specific phones in which the formed mouth shape is different from the mouth shape of the vowel. We defined these mouth shapes of the specific phones as "Beginning Mouth Shape" (BeMS). For example, a closed mouth which is formed when we utter "*ma*" or "*pa*" is one of the BeMS. We defined the mouth shapes that are same as the mouth shapes of the vowel as "End Mouth Shape" (EMS).

After that, we defined some codes for each mouth shape of the BaMS as "Mouth Shape Code" (hereinafter called MS Code). We also proposed an expression method of the BaMS using the MS Code. Because the BaMS are formed sequentially when we utter arbitrary words, we defined an expression of the sequence of the BaMS using the MS Code as "Mouth Shapes Sequence Code" (MSSC). Furthermore, we defined the mouth shape patterns of all Japanese phones using the MS Code. We call the mouth shape patterns of the MS Code "Phone Code". As the result, it is possible to convert words to the MSSC[5].

We consider that it is possible to read lips if the BaMS are detected from Japanese-speaking images. We propose a detection method of the BaMS from Japanese-speaking images. To detect them, we adopt template matching which assumes the BaMS to the template images.

2 Detection of the Basic Mouth Shape

As for the movement of the mouth shape, it is a repetition of changes from a certain BaMS to another BaMS when Japanese is uttered[6]. Similarity for each BaMS is calculated by the template matching. On the waveforms of the similarity, a partial waveform is flat in the term in which the EMS is formed, and a partial waveform forms convex in case of the BeMS[7]. These characteristics of the waveform are utilized to detect the BaMS. The BaMS is defined as (1), and each symbol expresses mouth shape /a/, /i/, /u/, /e/, /o/ and closed mouth, respectively. The BeMS is defined as (2), and the EMS is defined as (3)[5].

$$BaMS = \{A, I, U, E, O, X\} \tag{1}$$

$$BeMS = \{I, U, X\} \tag{2}$$

$$EMS = \{A, I, U, E, O, X\} \tag{3}$$

[1] The length of voice equivalent of one short syllable is called "mora", and the voice is called "phone".

2.1 Detection Method of the End Mouth Shape

As mentioned above, the partial waveform is flat during the EMS is formed. Here, the similarity of the n-th frame and the mouth shape $m(\in EMS)$ is expressed as $R(m,n)$. Therefore, if (4) is satisfied for $\forall m$, the n-th frame is assumed "EMS frame" (TH is a threshold of the similarity). For all frames speaking Japanese, it is determined whether each frame is the ESM frame or not.. Accordingly, it is able to divide an utterance term between the "EMS terms" and the others, and we call the latter term "BeMS term".

$$\left|\Delta R(m,n)\right| \le TH \tag{4}$$

$$\Delta R(m,n) = R(m,n) - R(m,n-1) \tag{5}$$

2.2 Detection Method of the Beginning Mouth Shape

The BeMS are not mouth shapes formed by all means unlike the EMS[2]. Therefore an idea is necessary to detect the BeMS. We know that a partical convex waveform is formed when the BeMS is formed from our previous study[7], and the mouth shapes are detected using this characteristic.

At first, each similarity is analyzed in the BeMS term. In a term $n_a \le n \le n_b (n_a < n_b)$ (see Fig. 1) in which (6) is satisfied, if (7) or $n_b - n_a > N_F$ is satisfied, the waveform in the term is considered as an upward slant to the right. Here, D and N_F are thresholds of the similarity and number of the frames, respectively.

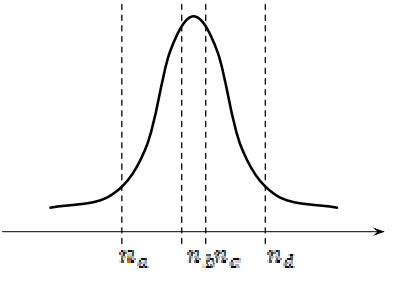

Fig. 1. The convex waveform which is formed when the BeMS is formed

$$\Delta R(m,n) > TH \tag{6}$$

$$\sum\nolimits_n \left|\Delta R(m,n)\right| \ge D \tag{7}$$

[2] A phone which is uttered without the BeMS is called "Simple Mouth Shape Phone", and a phone which is uttered with the BeMS is called "Couple Mouth Shapes Phone".

Next, in a term $n_c \le n \le n_d \left(n_b < n_c < n_d\right)$ in which (8) is satisfied, if (7) or $n_d - n_c > N_F$ is satisfied, the waveform in the term is considered as a downward-sloping.

$$\Delta R(m,n) < -TH \tag{8}$$

Finaly, if (4) and $n_c - n_b \le N_P$ are satisfied in a term $n_b \le n \le n_c - 1$, it is considered to be formed a BeMS between the n_a-th frame and the n_d-th frame from the waveform. Here, N_P is a threshold of number of the frames.

These processes are applied for each mouth shape of the BeMS, and peak value $R_P(m)$ is calculated by (9). The $R_P(m)$ is used for detection of the BeMS.

$$R_P(m) = \max\left(R(m,n_b), R(m,n_b+1), \cdots, R(m,n_c-1)\right) \tag{9}$$

3 Experiment for the Detection of the Basic Mouth Shape

To evaluate the method we proposed, we carry out some experiments to detect the BeMS and the EMS. The experiment contents are as follows:

1. Detecting the EMS from words that are composed by the Simple Mouth Shape Phone.
2. Detecting the BeMS from simple words that are composed by the Couple Mouth Shapes Phone.
3. Detecting the BaMS from words

Configuration of equipment carrying out the experiments is shown in Fig. 2. Images around mouth of a subject are taken with a digital video camera (hereinafter called DV camera), and the images are transferred to a PC.

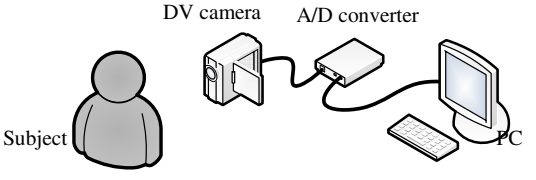

Fig. 2. Configuration of the experiments

During the experiments, distance between the DV camera and the mouth of subject are kept constantly, and the subject utters without moving head. Because the size of mouth is kept constantly and the horizontal and vertical movement of mouth is minimized. Before detecting the mouth shapes, the BaMS images which are used for the template are taken (see Fig. 3). In this experiment, the size of the frame images is assumed 640×480 pixels, and the size of the template images is trimmed to 495×375

pixels. When the template matching is performed, the images are converted into gray-scale images, and normalized cross-correlation is applied. The thresholds of constant to detect the BeMS and EMS are shown in Table 1. The first mouth shape and the last mouth shape of an utterance are assumed the closed mouth, and these two mouth shapes are not targeted for the detection.

Fig. 3. Template images

Table 1. The thresholds of constant

TH	D	N_F	N_P
0.02	0.06	3	4

3.1 Detection of the End Mouth Shape

The words used in this experiment are shown in Table 2. Each word is uttered five times. From the mouth shapes for which the averages of the similarity are greater than 0.6, up to two EMS having the highest and the second value are selected. In the selected BaMS, it is decided whether the EMS which should be detected is included or not.

Table 2. Test words and its MSSC to detect the EMS

#	Words (Japanese syllabary)	MSSC
1	ATSUGI	-A-U-I
2	KOPPU	-O-X-U
3	KOKESHI	-O-E-I
4	ENPITSU	-E-X-I-U
5	KONPAIRU	-O-X-A-I-U

The detection rates of EMS are shown in Table 3. The detection rates of each EMS which is included in the MSSC are shown in the column of "Detection rates", and the column of "As the highest similarity" shows the detection rates of the EMS which is detected as the highest average similarity. As a whole, the average detection rate of the EMS which was included in the two selected BaMS was 98.9%, and the rate which was detected the EMS as the highest similarity was 95.6%. We consider that these results are satisfactory. An example of the waveforms of the similarity is shown in Fig. 4. The horizontal axis of the chart shows the frame number, and the vertical axis shows the similarity R. The range of R is between -1.0 and 1.0, and 1.0 shows the best similarity. In addition, vertical lines in the chart separate the BeMS term and the EMS term. The ESM terms are shown with an arrow, and the previous terms are the BeMS term.

Table 3. Detection rates of the EMS

#	Words	Detection rates	As the highest similarity
1	ATSUGI	100.0%	100.0%
2	KOPPU	100.0%	100.0%
3	KOKESHI	100.0%	93.3%
4	ENPITSU	100.0%	95.0%
5	KONPAIRU	96.0%	92.0%
	Average	98.9%	95.6%

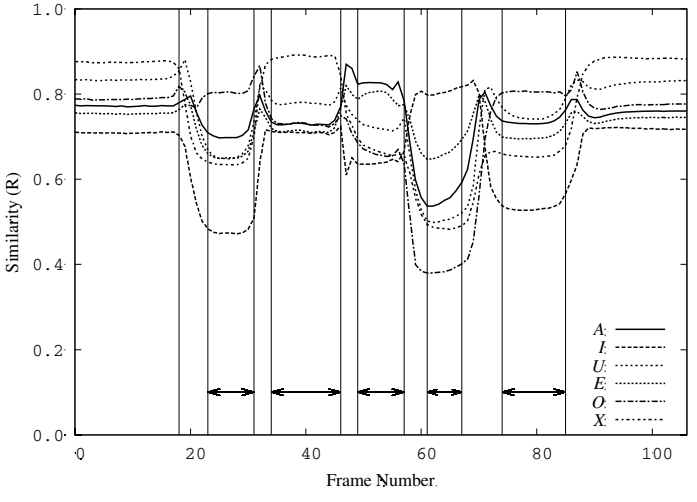

Fig. 4. Waveforms of the similarity about the test word "KONPAIRU"

3.2 Detection of the Beginning Mouth Shape

We choose simple words that are composed with two phones because detection of the BeMS is the purpose in this experiment. The test words are shown in Table 4. Same as the experiment of the EMS, the BeMS are detected from five utterances.

Table 4. Test words and its MSSC to detect the BeMS

#	Words (Japanese syllabary)	MSSC	#	Words (Japanese syllabary)	MSSC
1	ASA	-AIA	6	ASE	-AIE
2	NIWA	-IUA	7	UME	-UXE
3	UMA	-UXA	8	ASO	-AUO
4	KAMI	-AXI	9	IMO	-IXO
5	GAMU	-AXU			

The detection rates of the BeMS are shown in Table 5. As a whole, the detection rate of the BeMS was 93.3%, and the rate which was detected the BeMS as the highest peak value R_P was 88.9% (the column of "As the highest similarity"). We also consider that these results are satisfactory. An example of the chart is shown in Fig. 5. In the BaMS term which is shown with an arrow, the waveform of X forms convex. As a result, X is detected.

Table 5. Detection rates of the BeMS

#	Words	Detection rates	As the highest similarity
1	ASA	100.0%	100.0%
2	NIWA	100.0%	100.0%
3	UMA	100.0%	80.0%
4	KAMI	80.0%	80.0%
5	GAMU	100.0%	100.0%
6	ASE	100.0%	100.0%
7	UME	100.0%	80.0%
8	ASO	60.0%	60.0%
9	IMO	100.0%	100.0%
	Average	93.3%	88.9%

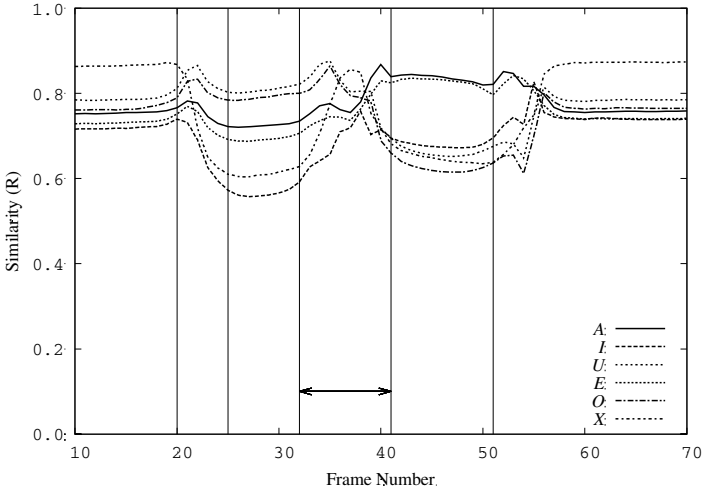

Fig. 5. Waveforms of the similarity about the test word "UMA"

3.3 Detection of the Basic Mouth Shape

The test words are shown in Table 6. Same as the previous experiments, the BaMS are detected from five utterances. Each detection rate is shown in Table 7. "Detection rates" shows the detection including the BeMS and the EMS, and the column of "False detection rates" shows the detection rate of the wrong BeMS which is not formed in the term. The detection rates were uneven, but the average detection rate of

the BaMS was 75.6%as a whole. The average detection rate of the BeMS was 67.3%, and the average detection rate of the EMS was 79.4%. These results are lower than the previous results. On the other hand, the false detection rate was 4.0%. Besides, the detection rate was 0.0% in some words. An example of the chart is shown in Fig. 6.

Table 6. Test words and its MSSC to detect the BaMS and the EMS

#	Words (Japanese syllabary)	MSSC
1	KATATSUMURI	-AIA-UXU-I
2	KAWAKUDARI	-AUA-UIA-I
3	KAMISHIBAI	-AXIXA-I
4	ASESUMENTO	-AIE-UXE-IUO
5	SUPOTTORAITO	-UXO-U-OIA-IUO

Table 7. Detection rates of the BaMS

#	Words	Detection rates	Detection rates of BeMS	Detection rates of EMS	False detection rates
1	KATATSUMURI	68.6%	100.0%	56.0%	13.3%
2	KAWAKUDARI	85.7%	50.0%	100.0%	0.0%
3	KAMISHIBAI	100.0%	100.0%	100.0%	0.0%
4	ASESUMENTO	68.9%	53.3%	76.7%	6.7%
5	SUPOTTORAITO	55.0%	33.3%	64.3%	0.0%
	Average	75.6%	67.3%	79.4%	4.0%

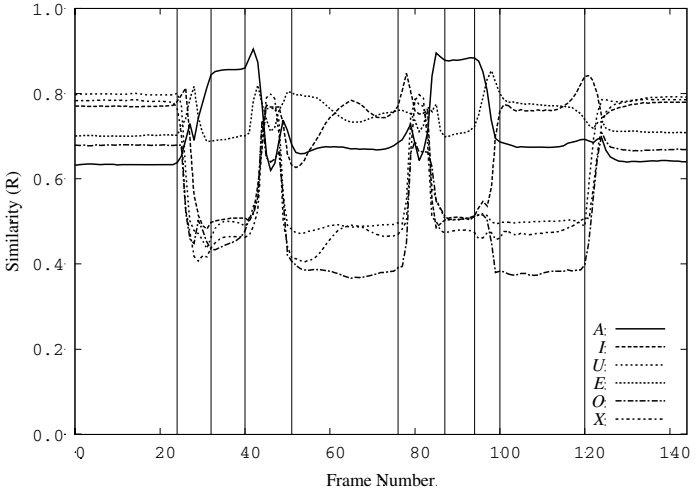

Fig. 6. Waveforms of the similarity about the test word “KAMISHIBAI”

We consider the cause that the detection rates of the BeMS and the EMS lowered. The underscored MS Code of the BaMS which have low detection rates are shown in Table 8. As a result, it may be said that the detection rates lower about I and U. However, all I and U are not low. As a result of analysis, we recognized two cases.

Table 8. The BaMS of low detection rate

#	Words	MSSC
1	KATATSUMURI	-AIA-UXU-I
2	KAWAKUDARI	-AUA-UIA-I
3	KAMISHIBAI	-AXIXA-I
4	ASESUMENTO	-AIE-UXE-IUO
5	SUPOTTORAITO	-UXO-U-OIA-IUO

Firstly, there are two mouth shape images about only U. One is an image in which teeth appear, the other is an image in which teeth do not appear. When we utter phones such as “SU”, “TSU” or “NU”, the teeth appear in the image. For this reason, we consider that the detection ratio falls.

Secondly, in the case the mouth shape of the BeMS is similar to the next mouth shape of the EMS, it is difficult to detect the BeMS. For example, it is difficult to detect U of the BeMS such as “SO” or “TO”, because the mouth shapes of U and O are similar. So, we have to consider the solution to these problems.

4 Conclusion

In this paper, we proposed a method which detects the BaMS from Japanese-speaking images. We adopt template matching to detect the BaMS, and the similarities were calculated. From the previous study, we confirmed that the waveforms of the similarity formed unique form when the BeMS and the EMS are formed. So, we paid attention to these characteristics and were able to detect the BaMS. We confirmed that the proposal method was effective from the experiments. At the same time, we also recognized some mouth shape patterns that were difficult to detect the BaMS. We have to review the solution to the problems, and it is necessary to improve the detection accuracy in future.

References

1. Kenji, M., Pentland, A.: Automatic Lip-reading by Optical-flow Analysis. The Transactions of the Institute of Electronics. Information and Communication Engineers J73-D-II(6), 796–803 (1990) (in Japanese)
2. Yasuyuki, N., Moritoshi, A.: Lipreading Method Using Color Extraction Method and Eigenspace Technique. The Transactions of the Institute of Electronics, Information and Communication Engineers J85-D-II(12), 1813–1822 (2002) (in Japanese)

3. Takeshi, S., Mitsugu, H., Ryosuke, K.: Analysis of Features for Efficient Japanese Vowel Recognition. The IEICE Transactions on Information and Systems E90-D(11), 1889–1891 (2007)
4. Kimiyasu, K., Keiichi, U.: An Uttered Word Recognition Using Lip Image Information. The Transactions of the Institute of Electronics, Information and Communication Engineers J76-D-II(3), 812–814 (1993) (in Japanese)
5. Tsuyoshi, M., Toyoshiro, N.: The Codification of Distinctive Mouth Shapes and the Expression Method of Data Concerning Changes in Mouth Shape when Uttering Japanese. IEEJ Transactions on Electronics, Information and Systems 129(12), 2108–2114 (2009) (in Japanese)
6. Tsuyoshi, M., Toyoshiro, N., Naohiro, I.: Evaluation for an Automatic Generation of Lips Movement Images Based on Mouth Shapes Sequence Code in Japanese Pronunciation. In: Proc. of Japan-Cambodia Joint Symposium on Information Systems and Communication Technology 2011 (JCAICT 2011) pp. 89–92 (2011)
7. Tsuyoshi, M., Toyoshiro, N.: A Deriving Method of Distinctive Mouth Shapes for Lip-reading. In: Multimedia, Distributed, Cooperative, and Mobile Symposium, pp. 1544–1549 (2009) (in Japanese)