Efficient Face Model for Lip Reading

Takeshi Saitoh

Kyushu Institute of Techonlogy, Fukuoka, Japan
saitoh@ces.kyutech.ac.jp

Abstract

There is number of researches on the lip reading. However, there is little discussion about which face model is effect for lip reading. This paper builds various face models which changes the combination of a face part, and changes the feature points. Various experiments were conducted on the conditions which change only model and do not change other algorithms. We apply the active appearance model. The CUAVE database which the utterance scene of ten digits is contained was used for the recognition experiment. As a result, the model which combined the external lip contour, nose, and outline acquired the highest recognition accuracy using HMM. We found that the face model which contains eyes and/or eyebrows is not effective for lip reading.
Index Terms: lip reading, word recognition, face models, active appearance model

1. Introduction

Speech recognition is widely used as an effective interface in many devices such as the personal computers, mobile phones, robots, and car navigators. In low-noise-level environments, an audio-only speech recognition (ASR) has obtained a high recognition accuracy and has been put to practical use. In noisy environments, however, the recognition accuracy degrades. Furthermore, a speech disordered people cannot use the ASR. It is difficult to use the ASR at the place where to utter voice is not desired, such as a public place.

Lip reading is one of the approaches which can expect to solve above problems, and the lip reading has attracted significant interest. There are two categories of lip reading, a pixel-based method which uses the information of the region around lip not needing accurate lip contour, and a model-based method which uses some shape features by extracting lip contour. The model-based method obtains the accurate recognition result by using lip contour.

Here, we have one question. Which face model is effective in lip reading? Some researchers may answer that the model of only a lip is good. Lip reading is a technique to recognize the contents of utterance. In order to utter, it is necessary to move a lip. Therefore, it is an appropriate answer that a lip is included in a model. However, there is no paper which analyzed this answer quantitatively.

It is an important fundamental research for lip reading to investigate the effective face model. In this paper, we build various face models. A lot of experiments were conducted on the conditions which change only face model and do not change other algorithms. We establish an effective face model for lip reading.

2. Conventional face models

This section surveys the face models conventionally used by lip reading. The number of points of a face model which are described below are summarized in Table 1.

Cox et al. and Hilder et al. build the face model not only including an external lip contour and an internal lip contour but eyebrows, eyes, a nose, and a face outline [2, 4]. Shin et al., Matthews et al., and Luettin et al. build the face model only by the lips of an external and internal [12, 6, 5]. However, the number of each feature points differs. In our previous research, our model consists of 16 points on the external lip, 12 points on the internal lip, and 10 points on the double nostrils, for a total of 38 points [10]. In [11], since the problem of a shooting angle, the model is close to the same of [10]. The number of nostril feature point differs. Active appearance model (AAM) is used by these all.

As the extraction method other than AAM, Nakamura et al. build the face model using an active contour model [7]. Their model includes the external lip contour, internal lip contour, nose, and outline. Nakamura et al. build the model only including the external lip contour with 16 points[8], and Deypir et al. build the model only 6 points [3].

Various face models are built by each research group. Although a difference can be confirmed with a recognition rate from reference, it does not discuss about the recognition accuracy by the difference in a face model. Since the recognition algorithm and experimental data are different, it is difficult to conclude which face model is effective in lip reading.

Table 1: Conventional face models and number of points.

reference	[4, 2]	[10]	[11]	[12]	[6]	[5]	[7]	[8]	[3]
outer lip	18	16	16	24	24	22	16	16	6
inner lip	16	12	12	22	20	16	16	–	–
eyes	12	–	–	–	–	–	–	–	–
eyebrows	8	–	–	–	–	–	–	–	–
nose	7	10	2	–	–	–	2	–	–
outline	12	–	–	–	–	–	5	–	–
total	73	38	30	46	44	38	39	16	6

3. Lip reading

3.1. Active appearance model

As described in [1], AAM is based on the idea of combining both shape and texture information about the objects that are to be modeled. The shape $\mathbf{x}$ is prepared for each of the N training images. These images are warped to a mean shape $\bar{\mathbf{x}}$ and normalized, yielding the texture $\mathbf{g}$. By applying principal component analysis (PCA) to the normalized data, linear models are obtained for both the shape, $\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s\mathbf{b}_s$, and the texture, $\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g\mathbf{b}_g$, where $\bar{\mathbf{x}}$ and, $\bar{\mathbf{g}}$ are the mean vectors, $\mathbf{P}_s$ and, $\mathbf{P}_g$ are the eigenvectors, and $\mathbf{b}_s$ and, $\mathbf{b}_g$ are sets of model parameters.

A given object can thus be described by with $\mathbf{b}_s$ and $\mathbf{b}_g$. As $\mathbf{P}_s$ and, $\mathbf{P}_g$ may still be correlated, PCA is applied once more using the following concatenated vector.

$$\mathbf{b} = \begin{pmatrix} \mathbf{W}_s\mathbf{b}_s \\ \mathbf{b}_g \end{pmatrix} = \begin{pmatrix} \mathbf{W}_s\mathbf{P}_s^T(\mathbf{x} - \bar{\mathbf{x}}) \\ \mathbf{P}_g^T(\mathbf{g} - \bar{\mathbf{g}}) \end{pmatrix}$$

where $\mathbf{W}_s$ is a diagonal scaling matrix derived from the value ranges of the eigenvalues of the shape and texture eigenspaces. This yields the final combined linear model $\mathbf{b} = \mathbf{P}_c\mathbf{c}$, where $\mathbf{P}_c = (\mathbf{P}_s^T, \mathbf{P}_g^T)^T$. In our method, the combined parameter obtained with the AAM is used as the recognition feature.

3.2. Algorithm summary

In our lip reading algorithm, we gives a lip position manually to an initial frame, and applies AAM using the position. For the subsequent frames, AAM is applied using the lip position detected with the previous frame. Here, we define various face models. The detail of the face models is described in **4**.

As mentioned above, the combined parameter obtained with the AAM is used as the recognition feature. A number of recognition methods have been proposed. In this study, we apply both DP matching and HMM which are well-known methods.

Figure 1: Original image of CUAVE database.

4. Experiments

4.1. CUAVE database

The CUAVE database [9] is a publicly available audio-visual database which contains speakers talking in frontal pose. This database consists of 36 speakers (19 male and 17 female speakers). This database is a speaker independent corpus and all the recorded speech is isolated-digits in English ('zero', 'one', .. ,'nine'). Each isolated-digit sequence was broken into the four tasks. For the purpose of this work, only the first normal task, where each speaker spoke 50 digits whilst standing still naturally, is used. The size of image is 720×480 pixels and its frame rate is 29.97fps. The background of each speaker is green as shown in Fig. 1.

The utterance scenes of 30 speakers (16 males and 14 females) among 36 speakers were used for this experiment. As for the number of the minimum frame, maximum frame, and average frame were 108, 3148, and 1581, respectively.

4.2. Experimental conditions

First, the utterance section of five times $\times$ ten digits was visually determined from movie. As for the result, the number of the minimum frame per word, maximum frame per word, and average frame per word were 10, 75, and 24, respectively.

In the following subsections, we described the experimental results with various face models. In these experiments, we used the leave-one-out method to obtain an accurate recognition rate with less data. That is, for each speaker, we divided the five samples into two groups of

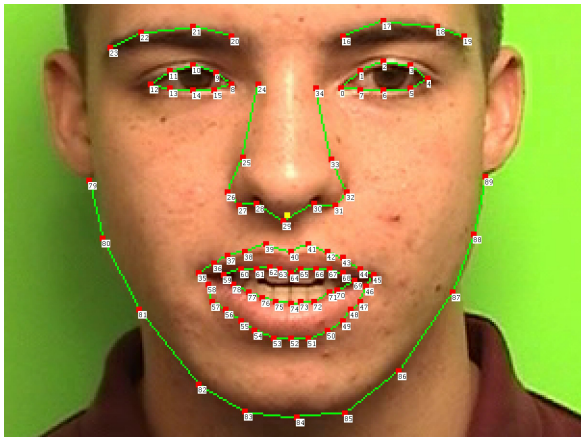

Figure 4: Feature points of base model M1.

four samples for training and one for recognizing. All experiments tested a speaker-dependent speech recognition, and the resulting average recognition rates with 30 speakers were computed.

We used an HMM toolkit (HTK). Here, the number of states of HMM was changed from 3 to 30, and a result with the highest recognition rate was selected as the result.

4.3. Facial parts

At first, the model which consists of eyes, eyebrows, a nose, an external lip contour, an internal lip contour, and a face outline as a base model M1 which includes the whole face as shown in Fig. 4 was built. The number of the external lip and the internal lip was set as 24 points and 20 points, these numbers are the same as [6]. The number in this figure is an index of feature point and the number of feature point of M1 is 90.

Next, nine derived models M2–M10 which combined the face part based on M1 were built. M2 is the model with which the model of only lips, M3–M6 are the models combined lips and other one part, and M7–M10 are the models combined lips and other two parts. Table 2 shows the details of each model, and Fig. 2 shows the mean shapes of each model. AAM was applied using ten built models and the combined features were computed.

The average recognition rates of all speakers of each model are shown in Fig. 3. In this graph, the red bars are recognition rates using DP matching, and the blue bars are recognition rates using HMM.

The average recognition rates of DP matching and HMM of the base model M1 were 65.2% and 69.8%, respectively. M5 which contains lips and the nose was obtained the highest recognition rate in four models (M3–M6) based on lips and other one part. M10 which contains lips, the nose, and the outline was obtained the highest recognition rate in four models (M7–M10) based on lips and other two parts.

4.4. Lip parts and number of feature points

The objective of previous experiment is to investigate the effective part and all models contained both the external lip contour and internal lip contour in the model. This experiment discusses the difference between the external and internal lip contour. M11–M14 were built as a derived model of M5 and M10. M11 and M13 are the models only in consideration of an external lip contour, and M12 and M14 are the models only in consideration of an internal lip contour. M5, M11, and M12 are models contained the nose which found the effective part for lip reading in the previous experiment. M10, M13, and M14 are models contained two parts of the nose and outline. The number of feature points of six models is shown in Table 2, and the average recognition rates are shown in Fig. 3.

In the previous experiments, the numbers of feature point of external lip contour and internal lip contour are 24 and 20, respectively. The objective of next experiment is the investigation of the number of feature points of lip contours. Three types of model were built. The number of feature points of M15 and M18 is the same as the previous study [10]. M16 and M19 have eight feature points of external lip and eight feature points of internal lip. M17 and M20 have four feature points of external lip and four feature points of internal lip. M5 and M15–M17 are models contained the nose. M10 and M18–M20 are models contained two parts of the nose and outline. The number of feature points of six models is shown in Table 2, and the average recognition rates are shown in Fig. 3.

From Fig. 3, M18 with HMM was obtained the highest recognition rate of 82.4%. The face model M18 consists of 16 points on the external lip contour, 12 points on the internal lip contour, 11 points on a nose, and 11 points on the outline. When analyzing the recognition rates of each speaker, 12 speakers among 30 speakers had acquired more than 90% of recognition rate. In this model, the standard deviation of 14.9% was obtained.

It was found that face parts effective for lip reading were a nose and a face outline except lips. The lip movement is not related the nose and the shape of the nose is always the same. It is guessed that a nose is a role of the reference part of a lips position. Moreover, since the motion of a face outline is the same as a motion of lips, especially the lower lip, the face outline is effective in recognition. On the other hand, a motion of eyes and eyebrows is independent of a motion of lips, and the model in consideration of these parts decrease the recognition accuracy. Interestingly, it became clear that the moderate number of feature point on the lip contour was better than too many feature points. However, the difference rate between M10 and M18 was less than 1%. This means that

Table 2: Twenty models and numbers of feature point.

model	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
external lip	24	24	24	24	24	24	24	24	24	24	24	–	24	–	16	8	4	16	8	4
internal lip	20	20	20	20	20	20	20	20	20	20	–	20	–	20	12	8	4	12	8	4
eyes	16	–	16	–	–	–	16	16	16	–	–	–	–	–	–	–	–	–	–	–
eyebrows	8	–	–	8	–	–	8	–	–	–	–	–	–	–	–	–	–	–	–	–
nose	11	–	–	–	11	–	–	11	–	11	11	11	11	11	11	11	11	11	11	11
outline	11	–	–	–	–	11	–	–	11	11	–	–	11	11	–	–	–	11	11	11
total	90	44	60	52	55	55	68	71	71	66	35	31	46	42	39	27	19	50	38	30

the number of feature point is not important problem.

5. Conclusion

This paper builds 20 face models which changes the combination of a face part, and changes the feature points. Various experiments were conducted to CUAVE database on the conditions which change only model and do not change other algorithms. As a result, the following knowledge was acquired.

○ The face model which contains lips, nose, and outline can obtain the highest recognition rate.

○ The face model which contains eyes and/or eyebrows is not effective for lip reading.

The contribution of this paper is establishment of the effective face model in lip reading.

6. Acknowledgements

The author thanks to Ms. Han Liang and Mr. Kenta Hara for assistance to this research.

7. References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *European Conference on Computer Vision*, number 2, pages 484–498, 1998.

[2] Stephen Cox, Richard Harvey, Yuxuan Lan, Jacob Newman, and Barry-John Theobald. The challenge of multispeaker lip-reading. In *Proc. of International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 179–184, 2008.

[3] Mahmood Deypir, Somayeh Alizadeh, Toktam Zoughi, and Reza Boostani. Boosting a multi-linear classifier with application to visual lip reading. *Expert Systems with Applications*, 38(1):941–948, Jan. 2011.

[4] Sarah Hilder, Richard Harvey, and Barry-John Theobald. Comparison of human and machine-based lip-reading. In *Proc. of International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 86–89, 2009.

[5] Juergen Luettin and Neil A. Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 1997.

[6] Iain Matthews, Timothy F. Cootes, J. Andrew Bangham, Stephen Cox, and Richard Harvey. Extraction of visual features for lipreading. *IEEE Trans. Pattern Anal. & Mach. Intell.*, 24(2):198–213, Feb. 2002.

[7] Kazuhiro Nakamura, Noriaki Murakami, Kazuyoshi Takagi, and Naofumi Takagi. A real-time lipreading lsi for word recognition. In *IEEE Asia-Pacific Conference on ASIC*, pages 303–306, 2002.

[8] Satoru Nakamura, Takao Kawamura, and Kazunori Sugahara. Vowel recognition system by lip-reading method using active contour models and its hardware realization. In *Proc. of International Joint Conference of SICE-ICASE*, pages 1143–1146, 2006.

[9] Eric K. Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N. Gowdy. Moving-talker, speaker-independent feature study, and baseline results using the cuave multimodal speech corpus. *EURASIP Journal on Applied Signal Processing*, 2002(11):1189–1201, 2002.

[10] Takeshi Saitoh. Development of communication support system using lip reading. In *Proc. of International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 117–122, 2011.

[11] Takeshi Saitoh and Ryosuke Konishi. A study of influence of word lip reading by change of frame rate. In *Proc. of International Conference on Auditory-Visual Speech Processing (AVSP)*, pages 131–136, 2010.

[12] Jongju Shin, Jin Lee, and Daijin Kim. Real-time lip reading system for isolated korean word recognition. *Pattern Recognition*, 44(3):559–571, Mar. 2011.

Figure 2: 20 face models.

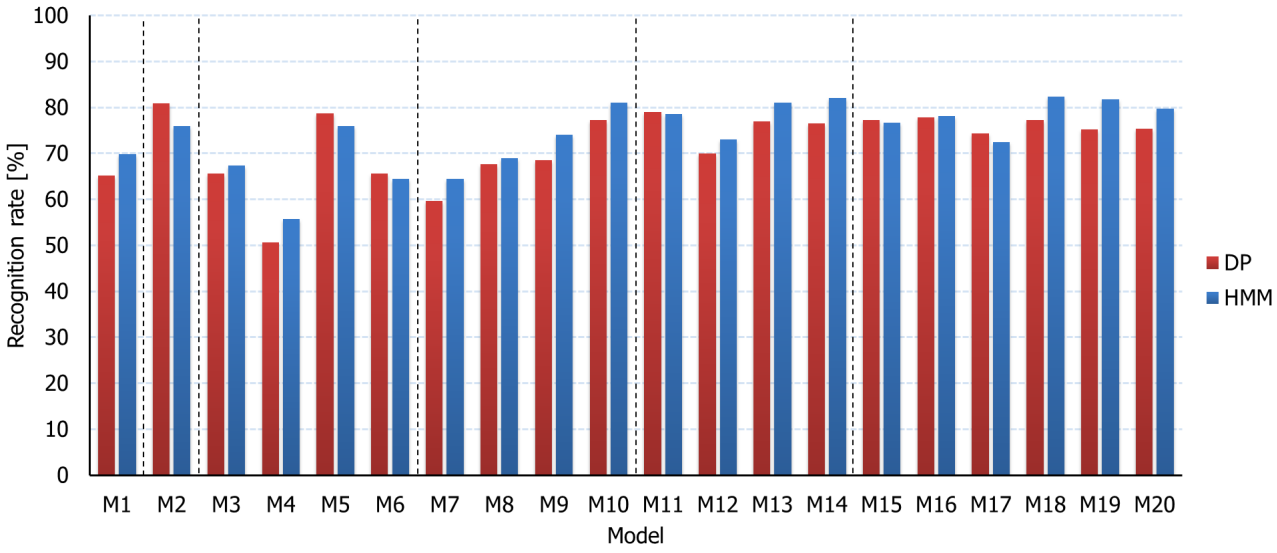

Figure 3: Recognition results.