A Study of Influence of Word Lip Reading by Change of Frame Rate

Takeshi Saitoh¹, Ryosuke Konishi²

¹Kyushu Institute of Technology, Japan
²Tottori University, Japan

Abstract

This paper discusses the influence of the word lip reading by change of the frame rate. The proposed method applies active appearance model to extract the face and several lip regions. Then, our method calculates trajectory feature and applies DP matching. We set the target words as the Japanese 25 words, and took 250 utterance scenes per a speaker from ten Japanese men. Though many of researchers on the lip reading using the utterance scene of 30fps, we took the utterance scene by 60fps. We changed the utterance scene of 60fps to the 11 kinds of pseudo utterance scenes. Using 12 kinds of frame rates, we carried out the recognition experiments with various combinations of the difference frame rate of the learning and recognition data, and evaluated the influence of the recognition accuracy. As a result, it was found that the frame rate of the recognition data is more sensitive than that of the learning data.

Index Terms: Word lip reading, trajectory feature, frame rate

1. Introduction

There have been a number of works concerned with lip reading. In this field, several topics, such as, word [1, 2, 3, 4, 5, 6] / single sound [7, 8], frontal face [1, 2, 5, 7, 8] / profile [3, 4, 5], Japanese [3, 6, 7, 8] / English [1, 2, 3, 4] / French [1], speaker dependent / independent speech recognition, only visual information [1, 5, 6, 7, 8] / integration of visual and audio information [2, 3, 4], are discussed. However, the research on a lip reading has few researchers with comparison research with speech recognition, and its recognition accuracy is not enough, either. On the other hand, as for lip reading technology, the use as an interface is expected like speech recognition technology. In recent years, some researchers are reported for real time lip reading technology [9]. In this case, it is assumed that not only the above-mentioned topics but the frame rate influences recognition accuracy. Then, in this paper, we investigate the influence of the lip reading by change of the frame rate. Here, we consider this research is basic problem of the lip reading, and we set a word as a recognition target and consider the photography direction as a front face and speaker dependent speech recognition.

The approach of the lip reading can be categorized as either appearance-based [1, 5], model-based [2, 3, 4, 6], or combination of the two. The appearance-based approach is based on the intensity and color information in a region of interest (ROI). This approach includes not only lip appearance but also information of the tooth and tongue. The dimensionality of the raw observation vector is often reduced using a linear transform. In contrast, the model-based approach assumes a top-down model of what is relevant for recognition, such as the lip contours. The parameters of the model fitted to the image are used as visual observations. Saitoh et al. proposed efficient feature called trajectory feature by model-based approach [6]. The lip reading algorithm in this paper is based on this literature [6].

2. Algorithm

2.1. Face extraction

The acquired image is shown in Fig. 1. Lip extraction is the most important process to obtain high recognition accuracy. A lot of methods for extracting the face and lip from such a facial image are proposed [10]. The active appearance model (AAM) proposed Cootes et al. [11] is a statistical model which the shape and grey-level appearance can be expressed by the low level for the object, such as, a face, an organ, which variously deformed from average shape in the image. Moreover, AAM can generate new image which is approximate the learning sample. AAM can extract two or more object regions, such as, the external and internal lip contour, simultaneously. This becomes possible to investigate an effective region for lip reading. For these reasons, we apply AAM to extract the lip regions. However, it is difficult to extract the target lip regions directly because several other parts, such as, eyes, eyebrow, and body, are reflected in the target image. Thus, we first extract the face region from the target image, and set a ROI to extract the lip region after the face extraction process. Here, we apply AAM to extract the face region as well as the lip regions.

To extract face region, we construct a face model. The face model consists of sixteen points on the right-and-left eye contour, ten points of the right-and-left eyebrow, and eleven points on the nose contour, total is 37 points as shown in Fig. 2. The purpose of not including lip contour in this model is to prevent the extraction error by lip motion in the utterance. Moreover, to be influenced from posture and the individual variation of the face, the facial contour is not included in this model. The extraction result to Fig. 1 is shown in Fig. 3. Here, the rectangle around lip is a ROI which describe in the next section.

2.2. Lip extraction

Next, we extract several lip regions. To extract these regions, we first set ROI around the lip to prevent wrong extraction. This ROI is square region, and the length of side is equals to the distance between right corner #0 and left corner #8 of the eye extracted from the face model as shown in Fig. 2. The location of ROI is based on the nose contour.

To extract lip regions, we construct a lip model. This model consists of sixteen points on the external lip contour, twelve points on the internal lip contour, and two points on double nostrils, total is 30 points as shown in Fig. 4. The reason for giving control point to the nostril is to prevent the extraction failure by the rapid lip motion during utterance. The nostril has little movement while uttering, and it can be stably extracted. Thus, the accuracy of the lip regions extraction can be stabilized.
Fig. 3. Face extraction result and ROI.

Fig. 2: Facial model.

Fig. 3: Face extraction result and ROI.

Fig. 4: Lip model.

Fig. 5: Lip extraction results.

Fig. 6: Five lip regions.

Fig. 7: Trajectory features.

2.3. Normalization of image space

Generally, speaker’s head moves when uttering. This movement is different depending on the individual. This movement influences the size and orientation of the face of our input image. Then, the size and orientation are normalized. The normalization process is as follows: The standard nose distance \( d_{0} \) is given, and the nose distance \( d_{n} = |#28 - #29| \) which are the number of the control point in Fig. 4, is calculated at each frame, and a scale ratio \( s_{n} = d_{n}/d_{0} \), is calculated. The orientation angle \( \theta_{n} \) between #28 and #29 is calculated. The affine transformation is applied using \( s_{n} \) and \( \theta_{n} \).

2.4. Shape features

Most of traditional methods used shape features based on the external lip region or intraoral region. On the other hand, we obtain five regions. Here, we define following thirteen shape features which are same as [6]: Four features of width \( W \), height \( H \), area \( A \) and aspect ratio \( A' \) are defined as 

\[
\begin{align*}
W &= \text{width}, \\
H &= \text{height}, \\
A &= \text{area}, \\
A' &= \text{aspect ratio}.
\end{align*}
\]

2.5. Trajectory feature

Saitoh et al. proposed a trajectory feature TF which obtain high recognition accuracy for lip reading [6]. TF is a time change of \( n \) features expressed as a \( n \)-dimensional trajectory of the lip motion of the target word. The trajectory is generated by plotting points by each frame features in \( n \)-dimensional space. Moreover, the feature points are interpolated by B-Spline curve.

TF is generated with several shape features. Here, the dimensions of each feature are differed. To avoid the influence of TF generation by this problem, we apply the normalization process. The average value of each feature may be set to 0 and variance may be set to 1 in the feature space.

Fig. 7 shows two TFS of word /a/me/ and /ko/n/ni/chi/wa/. In this figure, to generate TF (TF(S', A', S')) we use three shape features of \( S' \), \( A' \) and \( S' \). The blue points are the features of each frame, and the red curve is the interpolated curve.

2.6. Recognition method

In [6], the hidden Markov model and dynamic programming (DP) matching both are well-known method were applied, and both methods obtained enough high recognition accuracy. Then, this paper applied DP matching for recognition process.

Here, we denote \( TFX_{n} = \{x_{1}, x_{2}, \ldots, x_{j}\} \) is a target unknown TF, and \( TFX_{n} = \{r_{1}, r_{2}, \ldots, r_{j}\} \) is one of the reference TF. Then, the word \( \tilde{n} \) which satisfies an equation
The shortest word of $F$ was $w_{13} (/kyo/u)$, and the longest word was $w_{12} (/o/ka/e/xi/na/sa/1/)$. The difference between both numbers of utterance frames was 70, that is, 1.32 seconds. Moreover, the speaker E is the shortest utterance time of $T_E = 1.20$ seconds, and speaker B is the longest utterance time of $T_B = 2.18$ seconds. The time lag for 0.98 seconds arose to both. In this experiment, when speaking, we gave the speaker directions so that it might become the tone carried out clearly, but we have not given directions about utterance speed. The individual difference has appeared between utterance times.

### 3.2. Efficient feature

Before considering the influence of a frame rate, in TF, we discuss the efficient shape feature for lip reading. Based on computed thirteen shape features, we generated seventeen TFs. These combinations of TF were selected empirically. In this experiment, we applied leave-one-out method. The purpose of this research is to investigate the recognition accuracy for every speaker. We carried out the speaker dependent experiment.

The averaging recognition rates of ten speakers of each TF are shown in Table 3. As the results, the highest averaging recognition rate of 94.6% was obtained with TF($S^i, A^i, S^o$), and the second highest of 93.6% was obtained with TF($S^i, A^i, S^o$). These are exclusive relations although the difference of both combination is in $S^i$ and $S^o$. Although it is obvious, we can observe the tendency to obtain the recognition rate in which 3D TF is higher than 2D TF. Moreover, in 2D TF, TF($S^i, A^i$) is the highest and recognition accuracy becomes high further by including $S^o$ or $S^o$ in this. $S^i$ and $A^i$ are based on the lip region, and are unrelated to dental sight. On the other hand, since $S^o$ and $S^o$ are based on dental information, they can express the finer difference of a mouth shape.

The speaker J obtained 97.6% of the highest recognition rate and speaker D obtained 90.8% of the lowest recognition rate when TF($S^i, A^i, S^o$) was considered. When we pay attention the difference of the highest-lowest recognition rate between speakers, the combination of the smallest difference was TF($S^i, A^i, S^o$). That is, TF($S^i, A^i, S^o$) obtained a high recognition rate, and it became clear that this TF had a smallest change of the recognition rate by a speaker, and it found this TF is stable. We conclude TF($S^i, A^i, S^o$) is effective in recognition, and use this TF in next experiments.

### 3.3. Influence by frame rate

When using the lip reading technology as an interface, a real-time process is needed. Although the frame rate of a general video camera is almost 30fps, when using as an interface, we

### Table 1: Target 25 words.

<table>
<thead>
<tr>
<th>No.</th>
<th>word</th>
<th>number of frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>w01</td>
<td>/a/ri/ ga/to/u/</td>
<td>116</td>
</tr>
<tr>
<td>w02</td>
<td>/sa/yo/u/na/ka/</td>
<td>125</td>
</tr>
<tr>
<td>w03</td>
<td>/su/mi/ma/sa/ni/</td>
<td>102</td>
</tr>
<tr>
<td>w04</td>
<td>/ka/ma/na/la/</td>
<td>122</td>
</tr>
<tr>
<td>w05</td>
<td>/go/ku/ro/usa/ma/</td>
<td>143</td>
</tr>
<tr>
<td>w06</td>
<td>/o/ra/yo/u/</td>
<td>105</td>
</tr>
<tr>
<td>w07</td>
<td>/ko/n/ni/chi/ma/</td>
<td>135</td>
</tr>
<tr>
<td>w08</td>
<td>/ko/n/ba/n/ma/</td>
<td>130</td>
</tr>
<tr>
<td>w09</td>
<td>/o/ya/su/ni/ma/la/</td>
<td>160</td>
</tr>
<tr>
<td>w10</td>
<td>/ma/tar/ko/n/ko/</td>
<td>123</td>
</tr>
<tr>
<td>w11</td>
<td>/mo/shi/no/shi/</td>
<td>106</td>
</tr>
<tr>
<td>w12</td>
<td>/o/ka/e/xi/na/ma/la/</td>
<td>160</td>
</tr>
<tr>
<td>w13</td>
<td>/kyo/u/</td>
<td>81</td>
</tr>
<tr>
<td>w14</td>
<td>/a/shi/ta/</td>
<td>93</td>
</tr>
<tr>
<td>w15</td>
<td>/a/na/ta/</td>
<td>103</td>
</tr>
<tr>
<td>w16</td>
<td>/ki/na/ma/</td>
<td>95</td>
</tr>
<tr>
<td>w17</td>
<td>/o/to/to/1/</td>
<td>111</td>
</tr>
<tr>
<td>w18</td>
<td>/no/to/to/1/</td>
<td>80</td>
</tr>
<tr>
<td>w19</td>
<td>/a/me/</td>
<td>80</td>
</tr>
<tr>
<td>w20</td>
<td>/ku/mo/ra/</td>
<td>100</td>
</tr>
<tr>
<td>w21</td>
<td>/yu/xi/</td>
<td>86</td>
</tr>
<tr>
<td>w22</td>
<td>/a/tsu/1/</td>
<td>98</td>
</tr>
<tr>
<td>w23</td>
<td>/a/ta/ta/ka/la/</td>
<td>129</td>
</tr>
<tr>
<td>w24</td>
<td>/su/tsu/shi/la/</td>
<td>115</td>
</tr>
<tr>
<td>w25</td>
<td>/sa/mu/1/</td>
<td>101</td>
</tr>
</tbody>
</table>

### Table 2: Average number of utterance frame $F$ and utterance time $T$

| features | $A$ | $B$ | $C$ | $D$ | $E$ | $F$ | $G$ | $H$ | $I$ | $J$ | $K$ | $L$ | $M$ | $N$ | $T$ | $F_{avg}$ |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----------|
| $F_{frame}$ | 108 | 131 | 168 | 91 | 82 | 98 | 114 | 114 | 123 | 123 | 132 | 126 | 113 | 101 | 1.76 | 2.68 |
| $F_{sec}$ | 1.76 | 2.14 | 2.40 | 1.52 | 1.20 | 1.84 | 1.90 | 1.90 | 2.06 | 2.10 | 1.97 |
cannot always take an image by 30fps, and a frame rate changes with the performances of a lip reading algorithm or a computer. In this paper, we are taking the utterance scene by 60fps as described. Therefore, this experiment analyzes how a recognition rate is affected by influence in various frame rates.

It is assumed that the frame rates of learning and recognition are different. In this paper, the frame rate of the learning scene is denoted LFPS, and the frame rate of the recognition scene is denoted RFPS. We take the utterance scene by 60fps, therefore, we generated 11 kinds of pseudo frame rates of 30fps, 20fps, 15fps, 12fps, 10fps, 7.5fps, 6fps, 5fps, 4fps, 3fps, 2fps by 60fps, as shown in Fig. 9.

Using 12 kinds of frame rates which added 60fps, we changed LFPS and RFPS and carried out the recognition experiment. The result is shown in Fig. 10. Figure 10(a) shows transition of the recognition rate when changing LFPS, and Fig. 10(b) shows transition of the recognition rate when changing RFPS. When RFPS is almost the same as LFPS, the high recognition rate is obtained. On the other hand, the tendency for a recognition rate to fall when both difference becomes large is observed. The fall of a big curve is seen in Fig. 10(b). When computing the difference in height of the recognition rate by each condition, the large difference was seen in the condition of changing RFPS. It is assumed that RFPS influences to a recognition rate strongly from LFPS. Therefore, in order to obtain a high recognition rate, it is more desirable than LFPS to enlarge RFPS.

Moreover, when RFPS was the same as LFPS, and it was 10 or more fps, we obtained the recognition rate of not less than 90%. Generally, since learning data can be prepared in advance, computer use of high performance is expectable. Then, when it was assumed that we can take learning data by frame rate 30fps using a general camera, and recognition data could be taken by 15 or more fps, it became clear that we could obtain the recognition rate of not less than 90%.

3.4 Influence of the recognition rate by person

In order to analyze the details of a recognition result, the transition of average recognition rate of all the speakers at the time of changing LFPS in RFPS = 60 is shown in Fig. 11. The difference of all the speakers’ recognition rate is less than 10% at the time of LFPS ≥ 12. On the contrary, when LFPS is low, the difference of the recognition rate is large. At the time of LFPS = 4, the speaker J was 87.2% of the highest recognition rate, E was 34.8% of the lowest recognition rate, and both difference was 52.4%.

Next, we investigate the influence of the recognition rate by a speaker. The speakers’ E and J confusion matrix (CM) in RFPS = 60 and LFPS = 4 is shown in Fig. 12. CM contains information about actual and predicted word done by the recognition task. In this figure, the gray value of each cell shows the recognition rate. That is, the dark cell shows the high recognition rate. Conversely, the light cell shows the low recognition rate. The numerical value shows the recognition rate, and the recognition rate of 5% or more is written in these CMs.
Although the speaker E has five words (w01, w02, w06, w07, w09) of not less than 80% of a recognition rate from Fig. 12(a), there are eleven words (w03, w04, w05, w11, w14, w07, w09) of 20% or less of a recognition rate. On the other hand, the speakers J are 20 words and one word, respectively. That is, it is surmised that the recognition rate by the difference in a frame rate is affected by influence with a word.

In order to consider the above-mentioned surmise, we investigated the relation of the recognition rate between the utterance times \( T \) for every word. Figure 13 shows the relation of a recognition rate between the utterance time in all the speakers \( T_{dit}, T_{E} \) and \( T_{J} \). The conditions of Fig. 13(a) and (b) are \( L_{FPS} = 4, R_{FPS} = 60 \) and \( L_{FPS} = R_{FPS} = 60 \), respectively. \( T_{J} \) is longer than \( T_{E} \). The condition of \( L_{FPS} = 4 \) has few frames. There is a proportionality relation between \( F \) and \( T \), and it is obvious when two conditions (the utterance time is short and the frame rate is low) are satisfied, the number of frames decreases. It is a natural result that a difference arises to the speakers’ E and J average recognition rate. However, if the relation between a recognition rate and utterance time of the speaker J, the recognition rate cannot conclude easily as being affected by influence between utterance times.

Generally, if the vowel sequence is alike, incorrect recognition will take place easily. In order to check this, we calculated within-class variance between-class variance ratio \( J_{s} = \sigma_{W}^{2}/\sigma_{c}^{2} \) in the six mouth shape of Japanese five vowels and a closed-shape for every speaker. Here, \( \sigma_{W}^{2} \) is a within-class variance, and \( \sigma_{c}^{2} \) is a between-class variance. These variances are computed by the following equations.

\[
\sigma_{W}^{2} = \frac{1}{n} \sum_{i=1}^{c} \sum_{x \in X_{i}} (x - m_{i})^{2} (x - m_{i})
\]

\[
\sigma_{c}^{2} = \frac{1}{n} \sum_{i=1}^{c} n_{i} (m_{i} - m)^{2} (m_{i} - m)
\]

where, \( x \) denotes the feature vector, \( m_{i} \) denotes the average feature of class \( i \), \( c \) denotes the number of class, \( n_{i} \) denotes the number of sample of class \( i \), \( n \) denotes the total number of sample. \( J_{s} \) is large means that distributions of two classes is separated.

The numbers of mora of six mouth shapes (/a/, /i/, /u/, /e/, /o/, and /n/) of 25 words shown in Table 1, are 36, 21, 15, 5, 20, and 5, respectively. We obtained for the key frame which utters each mora from one every utterance scene of 25 words for every speaker visually. In particular, we select for the frame which appears vowel mouth shape visually, looking at mouth shape change of an utterance scene as shown in Fig. 9. Then, we calculated \( J_{s} \) between the six mouth shapes for every speaker. Table 4 shows \( J_{s} \) between six mouth shape of the speaker E and J. From this result, we can confirm that both tendencies are almost the same. This means that both difference for every mouth shape is almost the same. That is, we can consider that there is no difference in a clear tone, a natural tone, etc. arising from a speaker about the utterance scene used in this experiment.

From consideration to the above-mentioned, we assumed that it is not the difference in a tone that a difference arises to a recognition rate in a low frame rate. Under the influence of a frame rate, we think because the mouth shape used for recognition was not a suitable mouth shape which can express the contents of utterance.

4. Conclusion

This paper investigates the relation between a frame rate and recognition accuracy which seldom discuss until now. We set Japanese 25 words as the recognition target, and taken the utterance scenes with ten speakers. Moreover, based on the recognition result, we investigated the influence of the recognition rate by a speaker. As a result, we obtained the following knowledge.

- The most effective feature was \( TFi(S_{1}, A_{1}, S_{1}) \).
- When the frame rate of learning and recognition data is almost the same, we can obtain a high recognition rate.
On the other hand, if a difference arises in both frame rates, it will affect a recognition rate. However, in order to obtain a high recognition rate, it is desirable for the frame rate of recognition data to be higher than learning data.

- In order to obtain the recognition rate of not less than 90%, there are two photography conditions. (1) When $R_{FPS}$ is the same as $L_{FPS}$, the condition is they are 10 or more fps. (2) When $L_{FPS} = 30$, the condition is $R_{FPS} \geq 15$.

- In a low frame rate, not all the recognition rate of words falls equally, but is dependent on a word. However, the utterance time has not had big influence on the recognition rate.

- There was no big difference in the within-class variance between-class variance ratio between the mouth shapes by a speaker. It is not the difference in a tone that a difference arises to a recognition rate in a low frame rate. Under the influence of a frame rate, we think because the mouth shape used for recognition was not a suitable mouth shape which can express the contents of utterance.

In this paper, we carried out from the utterance scene taken in advance supposing using a lip reading as an interface. In order to prove the conclusion of this paper, the feature work is experimenting by making a lip reading algorithm into real time processing.

---

### Table 4: Variance ratio between each lip shape of speaker E and J.

<table>
<thead>
<tr>
<th></th>
<th>/a/</th>
<th>/i/</th>
<th>/u/</th>
<th>/e/</th>
<th>/o/</th>
<th>/n/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>0.65</td>
<td>0.73</td>
<td>0.16</td>
<td>0.99</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>/i/</td>
<td>1.48</td>
<td>0.58</td>
<td>1.15</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/u/</td>
<td>0.42</td>
<td>1.01</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/e/</td>
<td></td>
<td></td>
<td>0.30</td>
<td>1.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/o/</td>
<td></td>
<td></td>
<td></td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>/n/</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

### 5. Acknowledgements

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.

### 6. References


