Study of Differences in Individual Behavior Regarding Not-so-good Conditions and Corresponding Changes in Lip Motion while Speaking

Tsuyoshi Takahashi*a, Yoichi Kageyamaa, Masaki Ishiib, Makoto Nishidac, Etsuro Nakamuraa, Kenro Fujisawaa

aAkita University, 1-1 Tegata Gakuen-machi, Akita 010-8502, Japan
bAkita Prefectural University, 84-4, Aza Ebinokuchi Tsuchiya, Yurihonjo 015-0055, Japan
cThe Open University of Japan, 1-1 Tegata Gakuen-machi, Akita 010-8502, Japan

*Corresponding Author: tsuyoshi@ie.akita-u.ac.jp

Abstract

In our previous studies, we clarified that in case of humans, amusement can be detected by features that appear during lip-movement variation. Furthermore, we investigated the relationship between physical (health) condition and changes in lip motion of humans. The investigation revealed that the lip-motion width tended to fluctuate more widely in the horizontal direction. However, we have not yet investigated the differences in individual behavior regarding the relationship between physical conditions and changes in lip motion. In this study, we acquired lip-motion data of five persons and interviewed them about their physical condition over a period of 17 days. We analyzed the data to find the differences in individual behavior regarding lip-motion feature changes in cases when the subjects felt normal and did not feel good.

Keywords: lip motion, physical conditions, human sensing.

1. Introduction

Japan is the leading nation in terms of an aging population. Maintaining the health of the elderly is necessary for increasing social vitality. In order to realize this goal, the demand for health-monitoring systems using information technology, is increasing. A health-monitoring system requires a function to detect not-so-good conditions because early detection is important for sustaining the user’s health. The monitoring system must be simple to use and able to detect mild symptoms or physical fatigue in the user’s day-to-day health.

Lip movement while speaking is a type of behavioral feature, and can provide important information about the user. Some researchers have demonstrated the potential of using lip-motion features for command-input interfaces and lip-reading systems1–3. Lip-motion features are perceived as visual information; therefore, they are a natural basis for systems with non-contact-type interfaces. These systems are easy to use, and have advantages of high convenience and clean sanitary user operation. Additionally, lip-motion features can be acquired with or without vocalization, which gives them greater flexibility in situations that require silence. Our previous studies concerning the influence of vocalization on the command-input interface4 revealed that the lip motion of non-vocalized utterance differs slightly from that of vocalized utterance. According to reference4, the subject’s oral movements were exaggerated when uttering commands without vocalizing. Furthermore, the change in the lip motions due to the utterance experience of a word was less in non-vocalized utterance compared with that of vocalized utterance5.

Our previous studies6, 7 suggested that lip motions are affected by changes in a user’s psychological state or health condition. Specifically, we clarified that the lip-motion features based on height had the potential to reveal psychological changes when the subjects pronounced their names6. Furthermore, we clarified that amusement can be detected by lip-movement variations7. However, it may be difficult to detect amusement when the subject is not feeling well, e.g., conditions such as lack of sleep or catching a cold. Therefore, we investigated the relationship between a subject’s physical condition and changes in lip motion while speaking8. We acquired lip-motion data for five subjects; totaling to 3840 utterance data for all subjects combined. Further, we enquired about their physical and psychological condition for each day of the data acquisition.
Then, we performed a statistical analysis to find the differences in lip-motion features when the subjects felt normal and did not feel good. The results of the lip motion analysis suggested that the influence of physical condition tends to appear in the lateral movement of the lips. However, most data for a not-so-good-condition was acquired from a single subject in the previous study\(^8\). Therefore, the individual differences in the relationship regarding the physical conditions and changes in lip-motion have not yet been investigated in detail.

In this study, we acquired lip-motion data of five subjects (subjects A through E), and enquired about their physical condition over a period of 17 days. We performed an analysis to find the differences in individual behavior regarding lip-motion features when the subjects felt normal and did not feel good.

2. Data Used in the Experiment

2.1 Data Acquisition Environment and Flow

Fig. 1 shows the data acquisition environment and flow for an individual subject per day. We acquired utterance (speaking) video data for five subjects (subjects A–E; Japanese males and females in their twenties) uttering three Japanese words namely, “to/i/re,” “ro/u/ka,” and “yo/ku/shi/tsu.” The three words were selected based on our previous study\(^8\), considering operational interfaces of home-appliances using lip motions\(^1\). The subjects’ facial data were acquired from the front using a charge-coupled device (CCD) camera (Point Grey Research: Grasshopper; 640 × 480 pixel resolution; 60 frames per second (fps)). Table 1 lists the data acquisition conditions in this study.

The right side of Fig. 1 shows the data acquisition flow for an individual subject per day, using a three-step process. First, the subject sat in a chair in front of a monitor, remained quiet and rested, considering their own physical and mental condition. Second, the subject answered a questionnaire on their physical and psychological condition on the day of the experiment (we describe the details of the questionnaire in section 2.2). Third, the subject mouthed each utterance six times without vocalizing. We defined the six recorded utterance data as an “utterance dataset.” Then, 18 utterance videos (one utterance dataset per command) were recorded per subject per day.

The decision to have the subjects mouth each command six times without vocalization was based on the results of previous research\(^4\). These results revealed that the subject’s oral movements were exaggerated when uttering commands without vocalizing, and most subjects experienced negligible stress when uttering a single command up to six times.

Fig. 2 shows an outline of the data acquisition schedule for an individual subject. The first data acquisition period was scheduled for 17 days within a period of two months. A total of 306 utterance videos were recorded per subject. Thus, 1530 utterance videos were recorded for all three subjects combined.

The data used in this study were acquired with the approval of the ethical regulations regarding studies of humans at the Akita University, Japan.

2.2 Investigation Method for Physical and Psychological Conditions

In this study, we investigated the physical and psychological condition of each subject for each acquisition.
day. We prepared a survey form that included questions about the subject’s living, physical, and psychological conditions. Subjects were questioned using the same survey form in each acquisition day. Table 2 lists the questions for the questionnaire. The answers about physical condition were used for classifying utterance data; the acquired utterance data was classified as either normal-condition data or not-so-good-condition data based on the subject’s answers filled in the survey form.

We used the Affect Grid\(^9\) in the psychological condition survey to evaluate the subjects’ intuitive feelings on the day. The Affect Grid is a single-item scale with two dimensions as shown in Fig. 3. The vertical dimension represents the degree of arousal, with half the vertical dimension of the map representing a baseline feeling. The horizontal dimension represents the degree of pleasantness, with half the dimension of the map representing the baseline feeling, similar to the vertical dimension. In the part regarding psychological condition question, the subjects intuitively marked their existing feeling on the Affect Grid. The subject answered a questionnaire on their physical and psychological condition on the day of the experiment; the subjects marked their feelings before acquiring lip motion data. We calculated the average points of the arousal (A-Score) and pleasantness levels (P-Score) in normal-condition and not-so-good-condition data for each subject.

### Table 2. The questionnaire questions.

<table>
<thead>
<tr>
<th>Living condition</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bedtime on the previous day.</td>
<td></td>
</tr>
<tr>
<td>Today’s wake-up time.</td>
<td></td>
</tr>
<tr>
<td>Meal time before the experiment.</td>
<td></td>
</tr>
<tr>
<td>Exercise habits and frequency per week.</td>
<td></td>
</tr>
<tr>
<td>How long did you work yesterday?</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physical condition (Mark the applicable box)</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ Bad condition. (Terminate the exam).</td>
</tr>
<tr>
<td>□ Normal condition.</td>
</tr>
<tr>
<td>□ Not-so-good condition.</td>
</tr>
<tr>
<td>□ Mild symptoms caused by illness. (Please write down the symptoms).</td>
</tr>
<tr>
<td>□ Physical fatigue or muscle pain.</td>
</tr>
<tr>
<td>□ Other symptoms.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Psychological condition</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please fill in your mood. (Check the mark on the Affect Grid).</td>
<td></td>
</tr>
<tr>
<td>Please choose a word that is close to your current feelings.</td>
<td></td>
</tr>
<tr>
<td>□ Lively, □ Happy, □ Relaxed, □ Surprised, □ Fearful, □ Angry, □ Sleepy, □ Impatient, □ Sad, □ Heavy.</td>
<td></td>
</tr>
</tbody>
</table>

#### 2.3 Datasets

All the data were divided into “normal condition” and “not-so-good condition” data, based on the physical and psychological investigation performed at the time of the data acquisition. According to the results of the investigation, two symptom types were described in the subject’s answer sheet when they felt a not-so-good condition (i.e., a slight cold and physical fatigue). These two types of symptoms were used as not-so-good-condition data for analysis and experiment. The 1530 utterance data were divided into normal-condition data and not-so-good-condition data, based on the above definitions. Thereby, 1098 normal-condition data and 432 not-so-good-condition data were obtained. The breakdown of the not-so-good data was as follows: Subject A, 72 data (4 days’ worth of data); subject B, 144 data (8 days’ worth of data); subject C, 54 data (3 days’ worth of data); subject D, 108 data (6 days’ worth of data); subject E, 54 data (3 days’ worth of data).

### 3. Data Analysis

The preprocessing consisted of three steps as described in the following paragraphs.

#### 3.1 Preprocessing

Fig. 4 shows the outline of the lip-motion extraction process. First, the utterance videos were converted into static images (face images) at 60 fps, and an automatic fuzzy-reasoning method was used to extract the form of the lips\(^1\) for each frame to generate the data for analysis. Next, we manually extracted the utterance segments from the data. The operator determined the utterance segments in...
accordance with the criteria listed in Table 3. We needed to extract the utterance segments as accurately as possible, based on the criteria mentioned above. Therefore, automatic extraction methods for determining the utterance segments were not considered. Finally, we measured the horizontal length \( \text{di}X \) of the lip area in all the images. The lip area was defined by the rectangular region bounding the outline of the lip.

### 3.2 Computation of Lip-Motion Features

Motion changes in \( \text{di}X_n \) include small noises due to the light conditions or natural movement of the subject’s head during data acquisition. Therefore, the smoothing process proposed in our previous study was applied to remove the noise\(^6,7\). The smoothing process includes two steps namely, median filtering and averaging. First, each original \( \text{di}X_n \) data point is replaced by the median of the adjacent seven frames. Then, after an initial smoothing process, each data point is replaced with the local average of the adjacent data points. The window size of seven frames was adopted after testing a range of window sizes from three to nine frames.

In this process, two types of normalization were performed. First, we normalized the lip width \( \text{di}X_n \) to reduce the small changes in distance between the video camera and subject during data acquisition. A normalization process for \( \text{di}X_n \) was performed using Equation (1). The process used the lip size of the initial frame \( \text{di}X_1 \) and \( \text{di}Y_1 \) as the reference value. Then, the expansion degrees of the initial frame \( raX_n \) were calculated.

\[
raX_n = \frac{\text{di}X_n}{\text{di}X_1} \tag{1}
\]

Next, all the utterance segments were normalized. The speed of the lip motion and the total number of frames for each utterance were variable, even when the same subject uttered the same sentence. Therefore, all the utterance segments were normalized to a standard of 100 frames, and \( raX_i (raX \text{ of } i^{th} \text{ frame}) \) were linearly interpolated\(^{10}\) to obtain the normalized values \( X_i \).

In this study, we focused on the feature calculated from the lateral movement of the lip based on our previous study\(^{19}\). The lip motion feature used in this study is the average of the cumulative sum of the inter-frame difference in the width for each utterance dataset. The cumulative sum of the inter-frame difference \( DaX \) was calculated using Equation (2).

\[
DaX = \frac{1}{6} \sum_{j=1}^{6} \left( \sum_{i=1}^{F-1} |X_{ij} - X_{i(j+1)}| \right) \tag{2}
\]

Here, parameter \( X \) is the normalized value mentioned above, and \( F \) indicates the normalized frame number \( (F = 100) \). Furthermore, parameter \( i \) indicates the \( i^{th} \) frame of utterance data and parameter \( j \) indicates the amount of utterance data in a dataset. Parameter \( i \) ranges from 1 to \( F \) and \( j \) ranges from 1 to 6.

To analyze the changes in lateral lip movement, the analysis dataset of each subject was normalized using Equation (3).

\[
Z = \frac{(TF \cdot TF)}{\sigma} \tag{3}
\]

Here, \( TF \) indicates the target feature to be normalized \( (DaX) \), and \( Z \) indicates the normalized value of \( TF \). Parameter \( \sigma \) is the unbiased variance of the target feature.
4. Analysis Result

We investigated the differences in individual behavior regarding the lip-motion feature $DaX$ of the subjects for the normal and not-so-good condition. Table 4 summarizes the mean value of $DaX$ for all subjects. The $DaX$ values of the not-so-good condition data were different from the $DaX$ values of normal condition data for all the subjects. This shows that the physical condition affects the lip movement attributable to utterance. The difference in the value of $DaX$ between the normal and not-so-good condition was 0.74 points on an average for all subjects; “to/i/re” was 0.76 points, “ro/u/ka” was 0.68 points, and “yo/ku/shi/tsu” was 0.79 points. This tendency is particularly notable for subject E; the difference in value on an average for all words was 1.25 points. The minimum value was obtained for subject D; 0.26 points being the mean value. Next, when focused on the individual differences, the $DaX$ value for subjects A, B, D, and E was increased when they experienced a not-so-good condition. In contrast, the $DaX$ value of subject C showed a decreasing tendency for all the utterances. Table 5 shows threshold values for discriminating the normal and not-so-good condition data focused on $DaX$. The thresholds were set to obtain the maximum discrimination rate for each condition data based on previous study (8). The appropriate threshold differs for each subject as shown in Table 5. Most part of the not-so-good condition data of subjects A, B, D, and E distributed on upper side of the thresholds, and that of subject C distributed on lower side. These results revealed that the lip-motion feature $DaX$ included individual differences, and the differences were according to the amount of lip movement variation. Therefore, considering individual differences is necessary to develop a physical condition discrimination method using feature $DaX$.

5. Conclusion

In this study, we investigated differences in individual behavior of subjects regarding lip-motion changes attributable to the not-so-good condition. We acquired lip-motion data of five subjects and enquired about their physical condition over a period of 17 days. We performed an analysis to find the differences in individual behavior regarding lip-motion features when the subjects felt normal and did not feel good. Our experiments yielded the following conclusions:

- The lip-motion feature $DaX$, included individual differences, and the differences were according to the amount of lip movement variation.
- Considering differences in individual behavior, it is necessary to develop a physical condition discrimination method using feature $DaX$.

In future work, we plan to acquire more samples by increasing the number of subjects. We will also conduct further studies for developing the discrimination method.

Acknowledgment

The authors would like to thank Dr. C. Ishizawa from Akita University, Japan, for her assistance with the experiments; and Mr. Hiroki Ishibashi, Akita University, Japan, for his assistance in this study.

This study was supported by JSPS KAKENHI Grant Number 15K00222.

References