Fluctuation in Recognition Accuracy due to Utterance Experience in Command Input Interface using Lip Motion

Tsuyoshi Takahashi^{a,*}, Ayumu Saito^a, Yoichi Kageyama^a, Msaki Ishii^b, and Makoto Nishida^a

^aAkita University, 1-1 Tegata Gakuen-machi, Akita City, Akita 010-8502, Japan

^bAkita Prefectural University, 84-4 Tsuchiya-Ebinokuchi, Yurihonjo City, Akita 015-0055, Japan

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

Abstract

A command input interfaces based on lip motion features can also be used with or without vocalization. Our previous study revealed that the fluctuations of lip movement attributable to vocalization, and the difference of the command recognition accuracy of both utterance conditions. However, the change in lip motion due to utterance experience is not clear yet. To develop a practical command input interface using lip motion features, it is necessary to clarify the relationship between utterance experience and command recognition accuracy.

In this study, we acquired lip motion data of six participants for 10 days and conducted a command recognition experiment using DTW matching method to investigate the change in command recognition accuracy. According to our experimental results, command recognition accuracy decreased significantly when using vocalized data.

Keywords: lip motion, command recognition, utterance experience.

1. Introduction

Lip motion features due to utterances can provide important information for recognizing lip motions as commands. Because lip motion features are perceived visually, they are a natural basis for systems with non-contact-type interfaces. These systems have the advantages of high convenience and sanitation for users. Some researchers have demonstrated the potential of using lip motion features for command recognition⁽¹⁻³⁾. In our previous paper, we proposed a non-contact-type command input interface using lip motion features and reported that lip motion features such as changes in lip width and length

are useful for command recognition⁽³⁾. These are simple features that are easily acquired from sequential lip images and can be used to obtain high recognition accuracy. Additionally, these lip motion features can be acquired with or without vocalization, giving them greater flexibility in situations that may require silence. Unfortunately, the velocity and position of the lips may change under vocalization due to the action of other vocal organs⁽⁴⁾, and the number of frames varies depending on the speech speed. Thus, we investigated the fluctuations of lip movement attributable to vocalization. Our previous study revealed the differences in lip motion features between voiced data and non-voiced data and found that good recognition accuracy can be obtained in almost all cases when the input data and the registered data were of the same voicing state⁽⁵⁾. However, studies regarding changes in lip motion caused by utterance experience have not been conducted yet.

In this paper, we investigate the change in command recognition accuracy attributable to utterance experience. We acquired lip motion data while uttering five commands with and without vocalization for 10 days and conducted a command recognition experiment using the method given in reference (5). Our experiments led to some new findings for establishment of a command input interface using lip motion features.

2. Data Used

2.1 Acquisition Environment and Commands Used

In this study, we acquired sequential images obtained from six participants (id01–id06, all of Asian descent) as they uttered five commands. The data acquisition environment and process were essentially identical to the conditions given in reference (5). Figure 1 shows the data acquisition environment. Utterance data for the six

Fig. 1. Data acquisition environment.

participants were recorded by CCD video camera (Point Grey Research: Grasshopper) under the following conditions:

- The distance between the CCD video camera and the participant was fixed at 60 cm.
- A guide for lip position was shown to all participants. The dimensions of this guide were 75×40 pixels.
- Utterance videos were recorded under ordinary fluorescent lights (indoors, at a brightness of 650–800 lx).
- No lipstick or supplementary lighting was used.

The commands uttered were five Japanese words, each of which contained four to nine syllables. Table 1 lists the commands and the selection criteria used.

2.2 Data Acquisition Flow

Figure 2 shows an outline of the data acquisition schedule for an individual participant. Data acquisition was scheduled on 10 days within a period of three weeks for each respective participant. Figure 3 shows the data acquisition process per participant per day. Data acquisition was performed using a process similar to that in reference (5), involving four steps. First, the participant sat in the chair in front of the monitor and remained quiet and rested for 2 min to calm the participant's physical and mental state. Second, the participant uttered each command six times without vocalization. A 30 s interval was maintained between the groups of utterances of each distinct command. Third, participants rested for an additional 2 min to reduce the effect of utterances without vocalization. Finally, the participant uttered each command six times with vocalization as with the step 2. Sixty utterance videos were recorded per participant per day, totaling 600 utterance videos per participant over 10 days

Table 1. Commands used and the selection criteria.

Command	Number of syllables	Selection criteria
1: a/ki/ta/u/me/ko	6	- The word must contain all Japanese vowels.
2: o/o/da/te/ke/n/shi/ro/u	9	- The word must contain an "n" sound.
3: ka/ge/ya/ma/yo/u/i/chi	8	- The word must contain a plosive.
4: jo/u/ho/u/ko/u/ga/kka	8	- The word must contain a long vowel.
5: sya/shi/n/syu	4

Fig. 2. Data acquisition schedule for a single participant.

Fig. 3. Data acquisition process.

within a three-week period. Thus, 3600 utterance videos were recorded for all participants combined.

The decision to have participants utter each command six times was based on the results of a questionnaire used in previous research⁽⁶⁾. These results revealed that most participants experienced negligible stress when uttering one command up to six times.

3. Command Recognition Method

3.1 Preprocessing

Preprocessing was executed in four steps. First, the utterance videos were converted into sequential face images at 30 fps for acquisition of the lip motion features. Second, an automatic extraction method that employs fuzzy reasoning was performed to extract the shape of the lips⁽⁷⁾ in each sequential face image (see Figure 4). Third, we measured the horizontal length (diX) and vertical length (diY) of the lips in all lip images. Finally, we manually extracted the utterance section from each set of utterance data. The utterance section was determined by an operator in accordance with the following criteria:

- (a) The first frame of the utterance section was set to the frame immediately before the start of the lip movement.
- (b) The end frame of the utterance section was set to the frame immediately following the end of the lip movement.

Since we needed to extract the utterance section as accurately as possible using the criteria above, automatic extraction methods were not used.

3.2 Calculation of Lip Motion Features

Lip movements such as changes in lip width and length provide important information about uttered content and can be used for command input interfaces⁽³⁾. In this study, we conducted a command recognition experiment to investigate the variation of command recognition accuracy

$$raX_i = \frac{diX_i}{diX_1} \times 100 \quad (1)$$

$$raY_i = \frac{diY_i}{diY_1} \times 100 \quad (2)$$

$$R_i = \frac{diY_i}{diX_i} \times 100 \quad (3)$$

due to utterance experience in both utterance conditions (i.e., without vocalization and with vocalization) as described in reference (5). Three lip motion features (raX_i , raY_i , and R_i) were calculated in the same manner as in reference (3), using equations (1) to (3) below. Lip motion features raX_i and raY_i are the normalized values of diX_i and diY_i , respectively, and R_i is the ratio of lip width to lip length.

Fig. 4. Preprocessing.

Fig. 5. Process for generating standard patterns.

3.3 Generation of Standard Patterns

In this study, we generated standard patterns to use as the registered data based on the research results in references (3) and (5). Each standard pattern was generated from 12 utterances acquired on the first and second days. Figure 5 shows the outline of the three-step process for generating the standard pattern per command per voicing state. First, the lip motion features of 12 utterance data acquired on the first and second days were input. These 12 data have the same conditions (i.e., same command, same participant, and same voicing state) and thus create a unique pattern for each distinct combination of command, participant, and voicing state. Second, for a number of the utterance frames from the 12 data, the size was standardized in accordance with the mean size of the frames from the 12 utterances. Third, the mean values of each of the three lip motion features (raX_i , raY_i , and R_i) across the 12 utterance data were calculated, and these mean values were set as the feature values for the standard pattern. In this way, we created 10 standard patterns that included five non-vocalized and five vocalized patterns per participant; thus, 60 standard patterns (6 participants, 5 commands, 2

voicing states) were generated.

3.4 Data-matching Procedure

We calculated the similarity between the standard pattern and the input data using the inclination-limited DTW matching method used in references (3) and (5). The correspondence of matching nodes was limited to nearby two nodes, and weight coefficients were set to the same values as in the above references.

4. Experiment and Result

4.1 Outline of Experiment

Figure 6 shows the outline of this experiment. In order to consider the influence of utterance experience for

$$\text{Recognition rate} = \frac{\text{Number of successful data}}{\text{Total number of input data}} \times 100 \quad (4)$$

each utterance condition, we calculated the similarity between all combinations of input data for each day and the 10 kinds of standard pattern. Further, the matching experiment was performed on individual data; i.e., the same participant's input data and standard pattern were used. After the calculation, the pair having the minimum DTW score was then chosen for each command and voicing state. If the chosen pair was for the same command, it was determined that the command recognition was successful. The recognition rate was calculated using equation (4).

For the matching experiment, we used 2880 utterance data as input data that were acquired during the third day to the tenth day; the data consisted of 1440 non-vocalized data and 1440 vocalized data (i.e., six utterances per command and five commands per participant for each of the eight days from the third day to the tenth day).

4.2 Results and Discussion

Table 2 summarizes the experimental results using the same voicing condition for input data and registered data (i.e., matchups of input data to registered data were non-voiced data to non-voiced data, and voiced data to voiced data). The results show that good recognition rates of over 90% were obtained for both cases. On the other hand, focusing on the difference between voicing states, the overall recognition rate of non-vocalized data was slightly higher than that of vocalized data. This difference seems to increase as the acquisition day becomes later; a significant difference is seen after the 9th day. Figure 7 shows the

Fig. 6. Outline of command recognition experiment per participant.

Table 2. Results of experiment.

Input data	Recognition rate (%) (successful/total)	
	Non-vocalized data	Vocalized data
3rd day's data	95.6 (172/180)	93.3 (168/180)
4th day's data	98.3 (177/180)	94.4 (170/180)
5th day's data	95.0 (171/180)	93.3 (168/180)
6th day's data	93.3 (168/180)	95.0 (171/180)
7th day's data	97.2 (175/180)	91.7 (165/180)
8th day's data	95.6 (172/180)	92.8 (167/180)
9th day's data	92.8 (167/180)	83.9 (151/180)
10th day's data	97.2 (175/180)	90.0 (162/180)
All 8 days	95.6 (1377/1440)	91.8 (1322/1440)

Fig. 7. Change in command recognition rate.

change in command recognition rates for both voicing states over the eight days and the corresponding straight lines obtained by the least-squares method. The recognition rate has a decreasing tendency in both cases, with rates of decrease of 0.1% per day for the non-vocalized data and 1.0% per day for the vocalized data. For the non-vocalized data, command recognition accuracy decreased as participants became accustomed to uttering the commands, but the effect was small. On the other hand, command

recognition accuracy for the vocalized data decreased significantly as participants became accustomed to uttering the commands. In addition, the variance of the recognition rate across the eight days for the vocalized data was about 3.4 times that for the non-vocalized data. These results indicate an overall tendency for the characteristics of participants' mouth movement with and without vocalization: participants moved their mouth clearly when uttering commands without vocalization. The basic reason for the difference in mouth movement is due to awareness by participants at that time; when uttering commands without vocalization, they must pay attention to their mouth movements. On the other hand, when uttering commands with vocalization, they must pay attention to their voice. Thus, lip motion features are influenced by utterance experience when uttering commands with vocalization. It is revealed that the lip motion was changed due to utterance experience and lip motion fluctuation due to utterance experience in vocalized data was larger than that of non-vocalized data⁽⁸⁾. Therefore, it is necessary to develop an adaptive learning method for updating registered data to maintain high recognition accuracy.

5. Conclusions

This study investigated the relationship between utterance experience and command recognition accuracy. To investigate the influence of utterance experience, we acquired lip motion data for 10 days and conducted a command recognition experiment using the acquired data, calculating similarities between input data and the standard pattern for all combinations for each participant. Our experiments provide the following conclusions:

- Experimental results revealed that command recognition accuracy was slightly decreased in either voicing state: with vocalization or without vocalization.
- The command recognition accuracy of vocalized data decreased significantly as participants became accustomed to uttering those commands.

For future work, we plan to investigate the influence of utterance experience over a longer term. We also plan to develop an adaptive learning method for updating registered data to maintain high recognition accuracy.

Acknowledgment

The authors would like to thank Dr. C. Ishizawa, Akita University, Japan, for her assistance with the experiments.

This study has been supported by JSPS KAKENHI Grant Number 15K00222 and Grant Number 24500140.

References

- (1) T. Nakanishi, K. Terabayashi, and K. Umeda : "Mouth Motion Recognition for Intelligent Room Using DP Matching", IEEJ Trans. EIS, Vol. 129, No. 5, pp. 940-946, 2009
- (2) T. Saitoh and R. Konishi: "Real-Time Word Lip Reading System Based on Trajectory Feature", IEEJ Trans., Vol. 6, pp. 289-291, 2011
- (3) Y. Sato, Y. Kageyama, and M. Nishida: "Proposal of Non-Contact Type Interface of Command Input Using Lip Motion Features", IEEJ Trans. EIS, Vol. 129, No. 10, pp. 1865-1873, 2009
- (4) T. Sakai and K. Kawahara: Stricture Function and Materials of the Human body [Second Edition of Compact Edition], Japan Medical Journal, 2012
- (5) T. Takahashi, Y. Kageyama, A. Momose, M. Ishii, M. Nishida and A. Saito: "Basic Study of Lip Motion Fluctuation Due to Vocalization for Command Input Interfaces", IEEJ Trans. EIS, Vol. 134, No. 10, pp. 1429-1435, 2014
- (6) Y. Kageyama, Y. Ando, and M. Nishida: "Detection of Sentiments Changes Using Lip Motion Features of Utterance", IEEJ Trans. EIS, Vol. 131, No. 1, pp. 201-209, 2011
- (7) Y. Sato, J. Narita, Y. Kageyama, and M. Nishida: "Improvement of Automatic Lip Extraction Process Using Shape Information of Lips", IEEJ Trans. EIS, Vol.130, No.5, pp.873-881, 2010
- (8) T. Takahashi, A. Saito, Y. Kageyama, M. Ishii, and M. Nishida: "Analysis of Lip Movement Fluctuation Due to Utterance Experience", IEEJ Trans. EIS, Vol. 135, No. 10, 2015(accepted)