A Mobile Command Input Through Vowel Lip Shape Recognition

Yuto Koguchi, Kazuya Oharada, Yuki Takagi, Yoshiki Sawada,
Buntarou Shizuki, and Shin Takahashi

University of Tsukuba, Japan
{koguchi,oharada,takagi,sawada}@iplab.cs.tsukuba.ac.jp,
{shizuki,shin}@cs.tsukuba.ac.jp

Abstract. Most recent smartphones are controlled by touch screens, creating a need for hands-free input techniques. Voice is a simple means of input. However, this can be stressful in public spaces, and the recognition rate is low in noisy backgrounds. We propose a touch-free input technique using lip shapes. Vowels are detected by lip shape and used as commands. This creates a touch-free operation (like voice input) without actually requiring voice. We explored the recognition accuracies of each vowel of the Japanese moras. Vowels were identified with high accuracy by means of the characteristic lip shape.

Keywords: lip shape · vowel recognition · hands-free input · touch-free input · convolutional neural network

1 INTRODUCTION

Recent mobile devices depend greatly on touch operations. On the other hand, as touch screen sizes increase, it becomes difficult to operate the screen with only one hand. Therefore, in situations where the other hand is busy (for example, holding a briefcase), touch operations are difficult.

The use of hands-free inputs is a possible solution. Voice input is promising in this context. However, many people are uncomfortable when utterances in public spaces [1]. To solve this problem, 'voice' inputs without utterance (e.g., lip shape recognition) are being explored. For example, Lyons et al. [2] proposed a method of Japanese language input via lip shape recognition. This method does not require pronunciation, which is useful for situations where speech is to be avoided. It is, however, insufficient because vowels can be input via lip shape recognition but consonants must be entered by touch.

We propose a touch-free command input method for mobile devices using lip shape recognition. In our method, the lip shapes of vowels are recognized by the front camera of the device, and the sequence of recognized vowels are translated into commands. This allows for a touch-free input that is as intuitive as voice input but does not require the user to use his/her voice, avoiding the uncomfortableness from using it in public spaces.

2 RELATED WORK

We used lip reading to create touch-free input to a smartphone; words are guessed from lip shape. In this section, we review prior work on hands-free inputs and lip reading.

2.1 Hands-free Input Method

Several hands-free methods of operating electronic devices such as computers and smartphones have been studied. Orbits [3] allows operation of a smartwatch by moving the eyes in circles. The device identifies the direction and radius of eye rotation. For example, volume icons have been created whereby the volume can be changed (up or down) by moving the gaze clockwise or counterclockwise around the icon. CanalSense [4] of Ando et al. uses an earphone in which a barometer is embedded to measure the atmospheric pressure in the ear canal. This changes when the ear canal is deformed by movement of the jaw, face, or head. The changes are used to operate devices such as smartphones. The LUI [5] system uses lip shapes as commands to operate mobile devices. For example, when using a map, opening the mouth commands enlargement, and closing the mouth commands reduction. Compared to these methods, our method does not require additional devices, and users do not need to learn gestures or familiarize themselves with the application.

2.2 Lip Reading

Recent improvements in computer performance and advances in machine learning have rendered visual speech recognition increasingly accurate. Chung et al. [6] generated highly accurate subtitles from the mouth movements of a newsreader. Moreover, lip reading has been used to input commands to mobile devices. Lyons et al. [2] input Japanese-language commands via lip shape recognition. The consonants were entered manually and the vowels via lip shape. Compared to the usual Japanese input methods, this reduces the burden on fingers. Our method is based on real-time lip shape recognition of vowels, allowing for touch-free smartphone input with the lips only.

3 THE METHOD

3.1 Overview

We propose a method for control of a mobile device using lip shape. A user can perform shortcut operations by silently mouthing code words to the smartphone. Our method is shown in Fig. 1. The lip shape is extracted from the facial image of the front camera, and the vowel is recognized and input into the smartphone as a character. Finally, a shortcut command is estimated from the vowel sequence, and the action is performed. In the experimental section, we confirm the accuracy of our method.

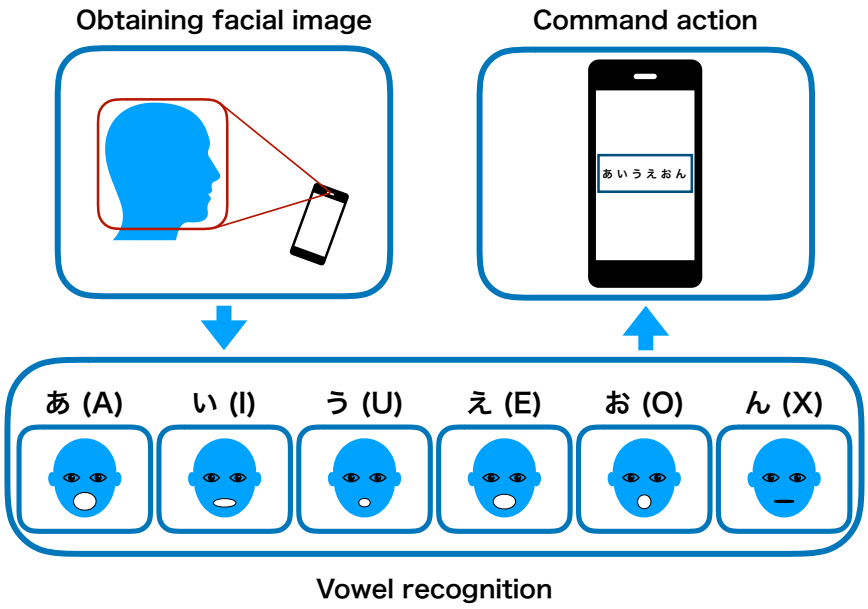

Fig. 1: Overview of our method.

3.2 Estimation of vowel sequence

In our method, lip shapes are recognized as vowels using the front camera. Then we use the mouth-shape code of Miyazaki et al. [7] to represent the vowels recognized from the lip shapes. The mouth shapes for word pronunciation are divided into six classes, composed of the five vowels (a, e, i, o, and u) and the closed mouth (x). In mouth-shape code, the initial lip shape (i, u, x) and the final lip shape (A, E, I, O, U, and X) for Japanese moras are combined to produce a code. More specifically, one mouth-shape code produced from a Japanese mora is a pair of one of the initial lip shapes (i, u, x) and one of the final lip shapes (A, E, I, O, U, X). For example, for the mora 'ma,' the initial lip shape is 'x' and the final lip shape is 'A,' producing the code 'xA.'

In this manner, changes in lip shape are recognized, and the mouth-shape code is stored as a 'vowel sequence.' For example, as shown in Fig. 2, the command 'Bu-ra-u-za,' which means 'Browser,' becomes the vowel sequence 'xU, iA, U, iA' after recognition and processing. Using such vowel sequences as inputs, commands can be executed.

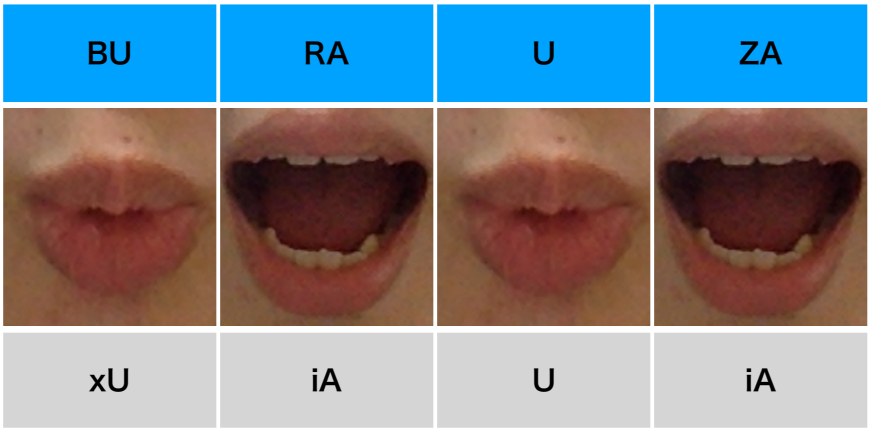

Fig. 2: Vowel sequence when 'Bu-ra-u-za' is spoken.

4 PROTOTYPE IMPLEMENTATION

We developed a prototype consisting of a smartphone and a server recognizing lip shape. Fig. 3 shows the system configuration. First, the user captures his/her face on the smartphone. Then, the smartphone sends the image to the server, which is implemented as an HTTP server. The server extracts a lip region from the received image, and returns a recognized vowel to the smartphone. If no face is found, the server returns 'none.'

We recognize vowels from lip shapes on facial images using a convolutional neural network (CNN). For the implementation of CNN, TensorFlow, an open-source machine-learning framework, was used. Fig. 4 shows the neural network configuration with nine convolutional layers. A pooling layer and a dropout layer are placed after the second, fourth, sixth, and eighth convolutional layers.

Six volunteers (aged 21–24 years; college or graduate students) were enrolled to provide their face images. We took facial images of all volunteers mouthing all vowels (and not mouthing at all) 100 times (3,600 images). We used the FaceLandmark Detector of Dlib (a machine-learning library) to extract the lips from each image. We augmented these data by adding 10 images generated by applying the following six random processes 10 times to each image:

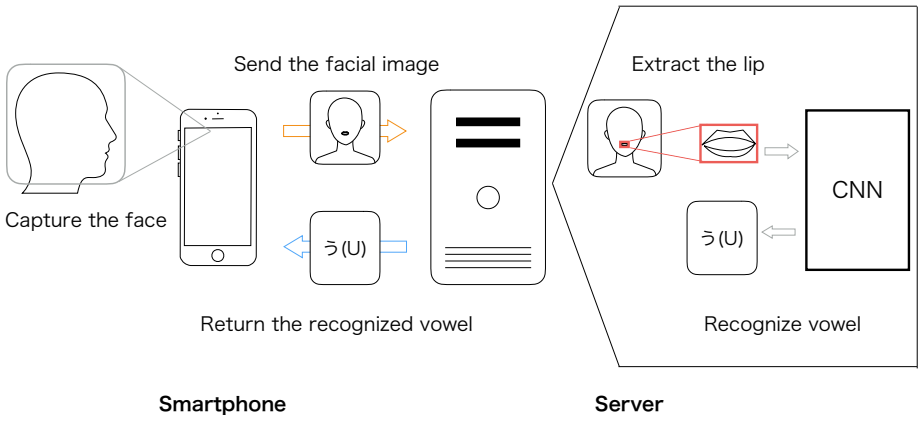

Fig. 3: System configuration.

- Rotate the image randomly from 0 degrees to 10 degrees.
- Shift in the horizontal direction within the range of 0% to 10% of image width
- Shift in the vertical direction within the range of 0% to 10% of image height
- Transform in oblique direction in the range of 0 to $\pi/8$
- Randomly shift the value of RGB in the range
- Randomly invert in the vertical direction

Therefore, we got 39600 images (3600 + 3600×10) as the learning data.

We created a learning model with a batch size of 50, 10 epochs, a multiclass logloss function, and an ADAM gradient. When 20% of the training data were used as test data, the accuracy was about 82%.

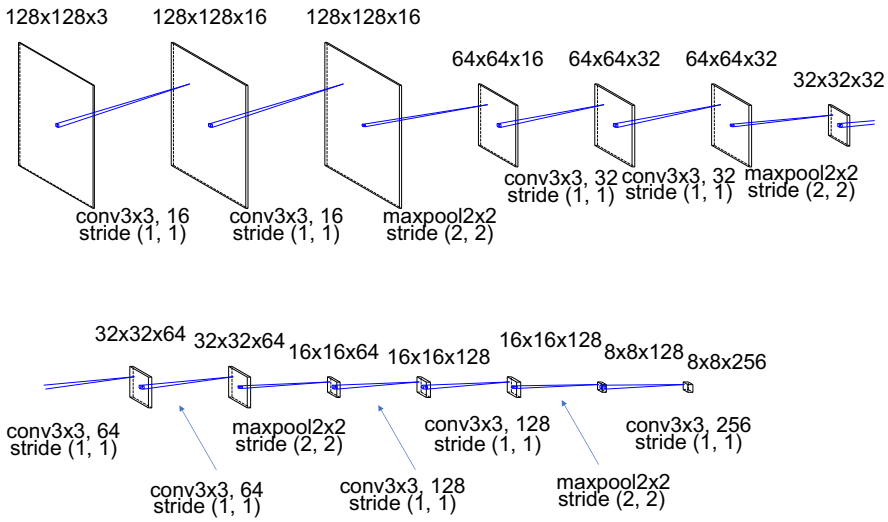

Fig. 4: The CNN model.

5 PRELIMINARY EXPERIMENT

We performed an experiment to explore the accuracy of our prototype system. We enrolled two volunteers (aged 22–23 years; college or graduate students) who had previously provided learning data and two volunteers (23–25 years; college or graduate students) who had not provided learning data. As the hardware of our prototype system, we used HUAWEI Mate 10 Pro for the smartphone and MacBook Air for the server.

5.1 Experiment Design

Each participant held a smartphone with the right hand at a distance of about 30 cm from the face.

First, the target mora was displayed on the screen (Fig. 5). Then the participant mouthed the target mora and simultaneously photographed his/her face with the front camera. Next, the target mora was refreshed and the process was repeated. This continued as long as the target mora was refreshed. We shuffled Japanese 75+1 moras (Fig. 6). They consist of 75 moras, which are the combinations of five vowels and 15 consonants, and the mora '𠮟 (X)'. Because

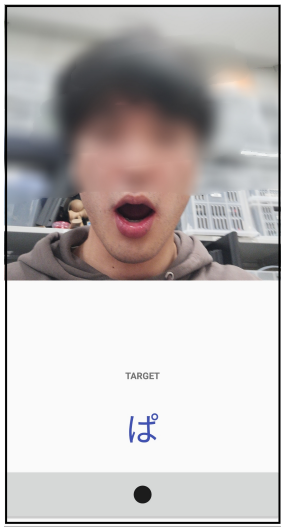

Fig. 5: Screenshot displaying the target mora ‘ぱ (PA)’.

we can obtain 15 images of each vowel from the 75 different moras, we made the mora ‘ん (X)’ appear 15 times as a target to collect the same number of images for the mora ‘ん (X)’. Thus, in total, each participant provided 90 facial images as test data. We measured lip shape recognition accuracy as the match rate between the target and recognized vowels in each task.

Prior to the experiment, we taught the participants how to use the prototype and allowed them to practice several times.

	-	k	s	t	n	h	m	y	r	w	g	z	d	b	p
A	あ	か	さ	た	な	は	ま	や	ら	わ	が	ざ	だ	ば	ぱ
I	い	き	し	ち	に	ひ	み	い	り	い	ぎ	じ	ぢ	び	ぴ
U	う	く	す	つ	ぬ	ふ	む	ゆ	る	う	ぐ	ず	づ	ぶ	ぷ
E	え	け	せ	て	ね	へ	め	え	れ	え	げ	ぜ	で	べ	ぺ
O	お	こ	そ	と	の	ほ	も	よ	ろ	を	ご	ぞ	ど	ぼ	ぽ

+

	-
X	ん

Fig. 6: The 75 + 1 target moras.

5.2 Result1: users who provided learning data

Table 1 shows the confusion matrix of our preliminary experiment for participants who had earlier provided learning data. Each rate is the proportion of the class in which the users' vowels were classified. The average recognition accuracy was 80.0%. In particular, 'あ (A),' 'い (I),' and 'ん (X)' were recognized with ≥90.0% accuracy. However, the recognition accuracy of 'え (E)' was only 53.3%, being often misrecognized as 'い (I).' Also, 'う (U)' and 'お (O)' were mutually misrecognized.

Table 1: Results for participants who provided learning data (%).

	あ(A)	い(I)	う(U)	え(E)	お(O)	ん(X)
あ(A)	96.7	3.3	0.0	0.0	0.0	0.0
い(I)	0.0	100.0	0.0	0.0	0.0	0.0
う(U)	0.0	0.0	63.3	0.0	36.7	0.0
え(E)	13.3	33.3	0.0	53.3	0.0	0.0
お(O)	0.0	0.0	20.0	3.3	76.7	0.0
ん(X)	0.0	0.0	10.0	0.0	0.0	90.0

5.3 Result2: users who did not provide learning data

Table 2 shows the confusion matrix of our preliminary experiment for participants who did not provide learning data. Each rate is the proportion of the class in which the users' vowels were classified. Characteristic mouth shapes such as 'あ (A),' 'う (U),' and 'ん (X)' were recognized with accuracies ≥80.0%. However, 'え (E)' was often miscategorized as other vowels, while 'お (O)' was misclassified as 'う (U).' Lastly, 'い (I)' was always misrecognized as 'ん (X).'

Table 2: Results for participants who did not provide learning data (%).

	あ(A)	い(I)	う(U)	え(E)	お(O)	ん(X)
あ(A)	80.0	0.0	0.0	16.7	0.0	3.3
い(I)	0.0	0.0	0.0	0.0	0.0	100.0
う(U)	0.0	0.0	96.7	0.0	0.0	3.3
え(E)	36.7	20.0	0.0	33.3	0.0	10.0
お(O)	0.0	0.0	36.7	0.0	46.7	16.7
ん(X)	0.0	0.0	0.0	3.3	0.0	96.7

5.4 Discussion

With new participants, 'ぃ (I)' was often mistakenly recognized as 'ん (X)'. Since the shape of these lips changes little, it can be considered that it was recognized as a closed state of the lips ('ん (X)') when mouthed the vowel 'ぃ (I)' moras. Also it can be considered that generalization performance of the CNN model was not high enough to accommodate individual differences in lip size and shape. As the volunteers who provided the learning data achieved high accuracy identification, we expect that the accuracy for regular users will improve as their own data are used to train the model.

In addition, when determining a command to be executed, there is a possibility of compensating unreliable vowel sequences by using string similarity metrics (e.g., the Levenstein distance). It is our future work to develop robust and reliable method of estimating commands from the recognized vowel sequences.

6 CONCLUSION AND FUTURE WORK

We proposed a method of operating a mobile device via lip shape, and conducted a preliminary experiment exploring the recognition rate. We used Dlib to extract the lip region and a CNN to recognize vowels. In a preliminary experiment, vowel recognition by characteristic lip shape was relatively accurate. We will improve the recognition rate by expanding the CNN model. We will also explore if users feel more comfortable with mouthing in public when comparing with voice input techniques.

7 ACKNOWLEDGEMENTS

We would like to thank Pedro Passos Couteiro for improving the English of the paper.

References

1. Sami Ronkainen, Jonna Häkkilä, Saana Kaleva, Ashley Colley, and Jukka Linjama. Tap Input As an Embedded Interaction Method for Mobile Devices. In *Proceedings of the 1st International Conference on Tangible and Embedded Interaction*, TEI '07, pp. 263–270, New York, NY, USA, 2007. ACM.
2. Michael J. Lyons, Chi-Ho Chan, and Nobuji Tetsutani. MouthType: Text Entry by Hand and Mouth. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, pp. 1383–1386, New York, NY, USA, 2004. ACM.
3. Augusto Esteves, Eduardo Velloso, Andreas Bulling, and Hans Gellersen. Orbits: Gaze Interaction for Smart Watches Using Smooth Pursuit Eye Movements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, UIST '15, pp. 457–466, New York, NY, USA, 2015. ACM.
4. Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. CanalSense: Face-Related Movement Recognition System Based on Sensing Air Pressure in Ear Canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, pp. 679–689, New York, NY, USA, 2017. ACM.

5. Maryam Azh and Shengdong Zhao. LUI: Lip in Multimodal Mobile GUI Interaction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pp. 551–554, New York, NY, USA, 2012. ACM.
6. Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Lip Reading Sentences in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3444–3453, July 2017.
7. Tsuyoshi Miyazaki and Toyoshiro Nakashima. The Codification of Distinctive Mouth Shapes and the Expression Method of Data Concerning Changes in Mouth Shape when Uttering Japanese. *The Transactions of the Institute of Electrical Engineers of Japan. C, A publication of Electronics, Information and System Society*, Vol. 129, No. 12, pp. 2108–2114, Dec. 2009.