The Seventeenth International Symposium on Artificial Life and Robotics 2012 (AROB 17th '12),
B-Con Plaza, Beppu, Oita, Japan, January 19-21, 2012

Producing text and speech from video images of lips movement photographed in speaking Japanese by using mouth shape sequence code - An experimental system to communicate with hearing impaired persons -

Shiori Kawahata, Eiko Koyama, Tsuyoshi Miyazaki[1], and Fujio Yamamoto[2]

Kanagawa Institute of Technology, Atsugi, Kanagawa 243-0292, Japan
[1]miyazaki@ic.kanagawa-it.ac.jp, [2]yamamoto@ic.kanagawa-it.ac.jp

Abstract: We proposed a method to detect distinctive mouth shapes from images that uttered Japanese. And a method to realize machine lip-reading from the order of detected mouth shapes was proposed. Therefore, we propose a communication system for hearing impaired persons using the machine lip-reading. This system supports the communication with the hearing impaired persons and remote persons using the Twitter. In addition, the input of the message is realized with machine lip-reading because hearing impaired persons have difficulty in utterance by the voice. We also devise the interface which can be operated only with a mouse. We carry out an experiment to send a message which is input with machine lip-reading to the Twitter, and evaluate this system.

Keywords: Hearing impaired persons, Twitter, Lip-reading, Communication

1 INTRODUCTION

For a hearing impaired person, both uttering and hearing a speech are difficult. Therefore, some special means are necessary to communicate with other people. As the one, there are lip-reading. The studies of so-called machine lip-reading using the image recognition technologies are performed. We are studying a method unlike the conventional machine lip-reading. We know that a person having the technology of lip-reading pays its attention to the distinctive mouth shapes and their transition, which appear intermittently during a speech. In our study, such knowledge from a human is used as the logical model for automatically recognizing the mouth shape movement. As the first stage, six kinds of basic mouth shape patterns are defined and expressed as the codes, which are easy to be handled by a computer. The pronunciation of Japanese words and sentences was proved to be expressed by series of such mouth shape codes. We called them MSSC (Mouth Shape Sequence Code) (Miyazaki et al [1]). In the second stage, we enabled the generation of MSSC from a video image, which is taken from the mouth movement of an uttering person (Miyazaki et al [2]). We consider that it is possible to generate a text string from utterance images using the machine lip-reading.

Acquired hearing impaired persons have ability to speak and to hear Japanese. This ability or experience is very useful for machine lipreading. In the hearing impaired persons, there are some persons that have physical disability, too. Therefore, we devise a system that sends a message to the Twitter from Japanese utterance images without using keyboard. We expect that this system is used as communication tools for hearing impaired persons.

In this paper, we describe the system that converts Japanese utterance images into a text message and sends the message to the Twitter. This system adopts the"mentions" to send messages to specific users. We propose the system that supports real-time communication with hearing impaired persons and remote place persons. The outline of this system is shown in Fig. 1.

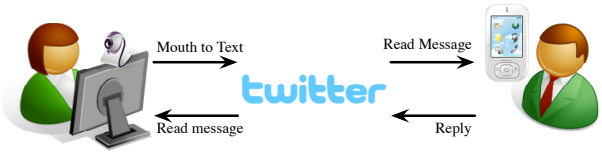

Fig. 1. Outline of this system

2 TWEET USING MACHINE LIP-READING

The block diagram of this system is shown in Fig. 2. This system consists of two subsystems (Part 1 and Part 2). The Part 1 generates MSSC mentioned above, from a video image (sound is not included), which photographed the movement of lips. It is sent to the Part 2 subsystem as text files. The Part 2 is controlled by a "Control window". By the operations on this Control window, Part 1 video picture is displayed, and the MSSC file is received. It is converted into the normal text. For this conversion, we have the pairs of MSSC code and the corresponding text in a phrases database. There is the case that different utterance contents are converted into a same MSSC code because of the principle of this conversion method. Therefore, it is necessary for the sender to choose an appropriate message among the candidates by some means. The chosen message is transmitted to the Twitter as a direct message. The reply text message from the addressee can be

confirmed with this GUI screen, too.

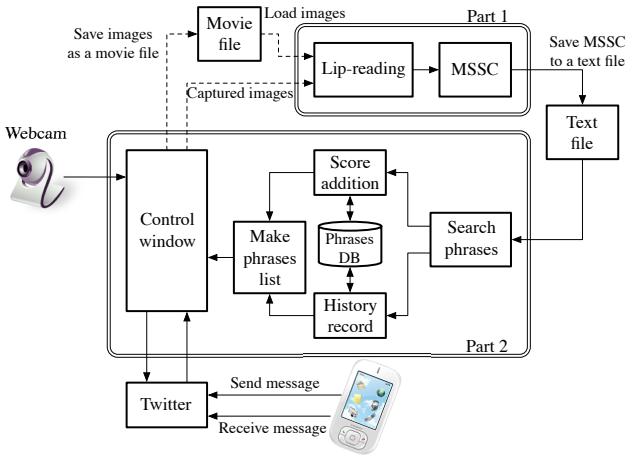

Fig. 2. Block diagram of this system

2.1 Generation of MSSC from utterance images (Part 1)

The lip-reading method which is used in this system is logically materialized knowledge of the lip-reading skill holders and detects distinctive mouth shapes. In the method, five mouth shapes of Japanese vowels and closed mouth shape were defined as "Basic Mouth Shape (BaMS)". We explain the relation between Japanese phones[1] and mouth shapes. A Japanese phone is uttered with only the "Beginning Mouth Shape (BeMS)" or the combination of the BeMS and "End Mouth Shape (EMS)". The BeMS are the mouth shapes which are formed at the beginning of utterance such as "ma", "sa" or "wa". The EMS are the mouth shapes which are formed at the end of utterance. When a Japanese words and phrases are uttered, the sequence of formed mouth shapes is expressed by "Mouth Shpaes Sequence Code (MSSC)". Each mouth shape is defined as $C = \{\mathtt{A}, \mathtt{I}, \mathtt{U}, \mathtt{E}, \mathtt{O}, \mathtt{X}, -\}$, which is called "Mouth Shape Code (MS Code)". Each MS Code corresponds to the mouth shape /a/, /i/, /u/, /e/, /o/ and closed mouth shape, respectively. The last MS Code "−" expresses that the BeMS is not formed. In an MSSC, the odd-numbered codes and the even-numbered codes correspond to MS Codes of BeMS and EMS respectively. For example, the MSSC of "EBINA" (name of a city in Japan) is expressed as "−EXI−A". This MSSC shows that a closed mouth shape "X" is formed on the second phone "BI".

In the Part 1, similarities between mouth shape image, which is captured by webcam or in movie file, and the BaMS images are calculated. After that, the BeMS terms and the EMS terms are detected from the utterance term using the similarities. Then, an MSSC is generated from the mouth shape of each term (Miyazaki et al [2]). Fig. 3 shows the generation process of MSSC.

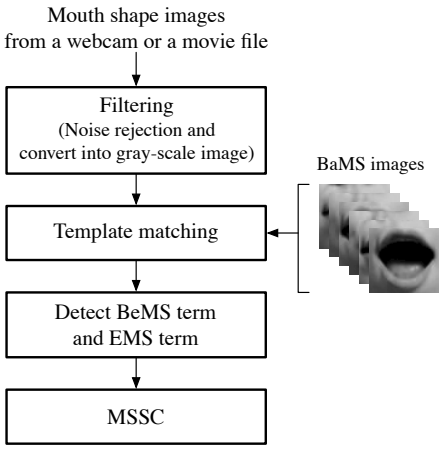

Fig. 3. The generation process of MSSC

2.2 Tweet from an MSSC (Part 2)

In this part, a Japanese phrase that is detected from the Phrases DB is sent to the Twitter (Makice [3]). To detect the phrase, two processes, "Score addition process" and "History record process", are executed. Table 1 shows the Japanese phrases and their MSSC stored in the DB.

After machine lip-reading, the phrase which the user required may not be detected. Therefore, in the Score addition process, points are added to each phrase stored in the DB.

In the History record process, a history count is added to the phrase which is sent to the Twitter. As the result of the two process, some phrases are listed by the score order and the history order. Then, the user selects a phrase from the list, and the phrase is sent to the Twitter.

2.2.1 Score addition process

To add a score to each phrase stored in the DB, each MS Code of MSSC generated in the Part1 is compared with the MS Code of the phrases stored in the DB. The score addition flow is shown in Fig. 4. In the Fig. 4, $c_n(s)$ means a s-th MS Code of MSSC of n-th phrase stored in the DB. $m(s)$ means a s-th MS Code of the MSSC generated in the Part 1 correspondingly.

An example is shown in Table 2. This result shows the case of utterance "Arigato". In the MSSC row, the MS Codes of "Arigato" stored in the DB are shown. In the Result row, the MS Codes that are generated in the Part 1 are shown. The row of *score*1 and *score*2 show the points that are calculated in the Score addition process. For example, the point of *score*1 in the case of $s = 2$ is 1 because of $c_n(2) = m(2)$. The point of *score*1 in the case of $s = 6$ is 0 because of $c_n(6) \neq m(6)$.

[1]The length of voice equivalent of one short syllable is called "mora", and the voice is called "phone".

Table 1. Phrases Database

#	MSSC	Japanese articulation	English meaning
1	-I-U-A-E-IXA-U-A	Itsu kaeri masuka	When do you come home?
2	-IXAUOIE-U-A	Ima doko desuka	Where are you now?
3	-I-X-U-I-A-I-E-U	Kinkyu jitai desu	State of emergency.
4	-A-I-AUO-U	Arigato	Thank you.
5	UA-IXA-I-A	Wakari masita	I've got it.

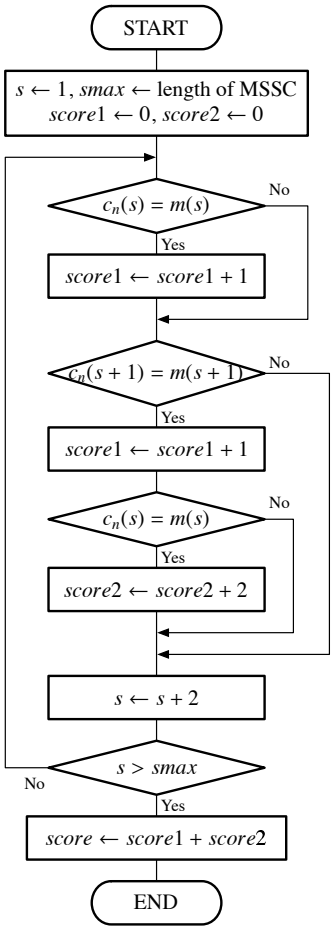

Fig. 4. The flow of Score addition process

About the *score*2, the first point is 2 because of $c_n(1) = m(1)$ and $c_n(2) = m(2)$. The third point is 0 because of $c_n(5) = m(5)$ and $c_n(6) \neq m(6)$.

Then, the total of *score*1 becomes 7 and the total of *score*2 becomes 4. As a result, the score of n-th phrase becomes $score1 + score2$.

2.2.2 History record process

In this process, the count that a phrase is sent to the Twitter is recorded. If some phrases get the same score, the phrase of which the count is greater is listed higher on the Control window. The phrases that are used frequently become easy to be selected.

Table 2. Example of point addition

s	1	2	3	4	5	6	7	8	9	10
MSSC	-	A	-	I	-	A	U	O	-	U
Result	-	A	-	I	-	E	-	O	-	O
*score*1	1	1	1	1	1	0	0	1	1	0
*score*2	2		2		0		0		0	

Firstly, The phrase list is sorted by the history order. Secondly, the list is sorted by the score order and some of top phrases of the list are displayed.

2.3 Control window

We describe the Control window of this system. The window is shown in Fig. 5. Four components are located in this window.

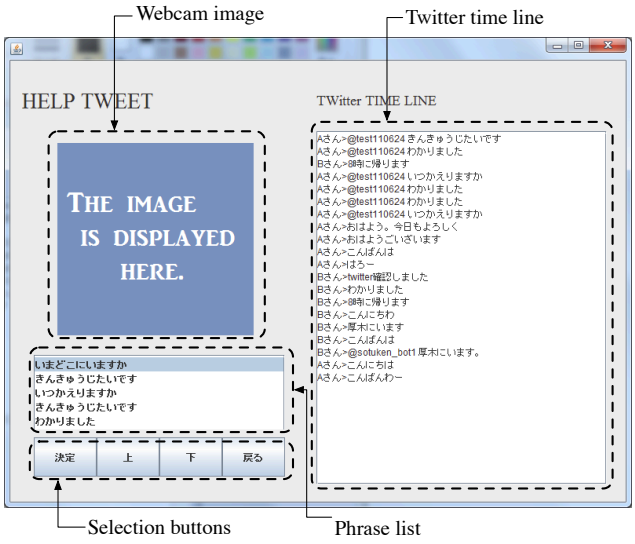

Fig. 5. Control window

The images captured by webcam are shown in the "Webcam image" pane. As the result of machine lip-reading, top five phrases are listed on the "Phrase list" pane. The user select a phrase from the list by using selection button and send a message to the Twitter. The four buttons are "Send", "Upward", "Downward" and "Cancel". After that, the posted message is listed on the top of the "Twitter time line".

Table 3. The detection rate and the ranking

Japanese articulation	English meaning	1st	2nd	3rd	4th	5th
Ashita	tomorrow	83% (1)	100% (1)	83% (1)	100% (1)	33% (2)
Ima	now	75% (1)	50% (1)	75% (1)	75% (1)	75% (1)
Istu	when	75% (1)	50% (2)	75% (1)	50% (2)	100% (1)

3 EXPERIMENTS

We carried out an experiment to convert into Japanese words and phrases from MSSC. The MSSC were generated from images that uttered Japanese. However, we used recorded utterance images.

Firstly, Japanese phrases and their MSSC were stored in the Phrases DB. The stored phrases were "Kyo" (today in English), "Ashita" (tomorrow), "Ima" (now), "Itsu" (when), "Dokoni" (where), "Dokohe" (where), "Imasuka" and "Kaerimasuka".

During the experiments, the subject uttered phrases without moving head. Furthermore, the mouth shapes before utterance and after utterance assumed closed mouth shape.

3.1 Experimental results

A subject uttered Japanese phrases five times. The phrases were "Ashita", "Ima" and "Itsu".

The detection rate of each phrase and its ranking are shown in Table 3. The detection rate shows the ratio that is matching rate of each MSSC stored in the DB and generated in the Part 1. The number in parenthesis shows the ranking.

In most results, the detection rate was higher than 50%, and some of them got 100%. Besides, all phrases were ranked as the 1st or 2nd in the list. Through this experiment, we confirmed that the detection rates differed, though the same phrases were uttered.

4 CONCLUSION

In this paper, we proposed a system for supporting hearing impaired person by using machine lip-reading. An MSSC was generated from the utterance images and was able to be converted into Japanese. Furthermore, the converted Japanese was able to be sent to the Twitter as a message.

If the phrases in the Phrases Database are increased, incorrect phrases may be listed in high-ranking. As a remedy, we have to devise score addition method and examine comparison method of MSSC.

Although the system is only at a very early stage now, we can see that this approach has big possibilities to promote communication between hearing impaired persons and the general people. Additional facilities such as text-to-speech conversion and speech-to-text conversion are also being developed in order to provide more convenient usages. We also plan to improve the precision of the conversion from mouth shape images to readable text.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology Grant-in-Aid for Scientific Research (C) 23501176 and Grant-in-Aid for Young Scientists (B) 23700672.

REFERENCES

[1] Miyazaki T, Nakashima T, Ishii N (2011), A Proposal of Mouth Shapes Sequence Code for Japanese Pronunciation. The 12th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2011), Studies in Computational Intelligence 368, Sydney, Australia, Jul 6-8, 2011, pp. 55-65

[2] Miyazaki T, Nakashima T, Ishii N (2011), A Detection Method of Basic Mouth Shapes from Japanese Utterance Images. Proceedings of the 14th International Conference on Human-Computer Interaction (HCII 2011), Lecture Notes in Computer Science 6761, Orlando, Florida, USA, Jul 9-14, 201, pp. 608-617

[3] Makice K (2009), Twitter API: Up and Running Learn How to Build Applications with the Twitter API. O'Reilly