# A Novel Transducer: From Lip Motion to Voice Message

Takeshi Saitoh, Tomoya Kato and Ryosuke Konishi
Tottori University
4-101 Koyama-minami, Tottori, 680-8552, JAPAN
e-mail {saitoh, konishi}@ele.tottori-u.ac.jp

## Abstract

*This paper describes the novel transducer which recognizes lip motion of the word utterance and output the recognized word as the voice message. Though, the base method of our transducer is speaker dependent word lip reading, our system is adopted the facial movement of the user. For the developed prototype system, the robustness was evaluated to the two kinds of facial movement, the distance between the user and camera changed and rotation movement of the face. As the result, we confirmed the robustness above movements. Moreover, the recognition experiment with three subjects was carried out for two word groups of 14 words and 47 words. The recognition rates of 90.0% and 82.6% were obtained with 14 words group and 47 words group, respectively.*

## 1 Introduction

There are a number of speech handicapped person. Our aim is to development of the communication support system for such person. As the first step, this paper describes the speaker dependent novel transducer which recognizes the lip motion of the word utterance and output the recognized word as the voice message.

Speech is the most natural communication means for human. However, this means is difficult for the speech handicapped person who can not utter clear voice. Their main means are by writing or sign language, and these means takes burden with both the speaker and listener. Then, we focus the lip reading which recognizes the utterance meaning based on visual lip motion. But the level of lip reading is low compared with the speech recognition based on auditory information [1], and there is few research of the application with lip reading. Though, it is desirable to recognize the conversation sentence, there is no report about sentence recognition in the field of lip reading. Therefore, in this paper, we focus not the sentence recognition but the word recognition, and develop the application to recognize the word in real time and output the voice message by the voice synthesis.

Concerning the application of lip reading, several researches were proposed by using not the lip motion but the still mouth shape in real time [2, 3, 4]. Lyons et al. proposed the character input system that integrated key operation and five vowel mouth shapes [2]. Their system is based on semi automatic recognition method,

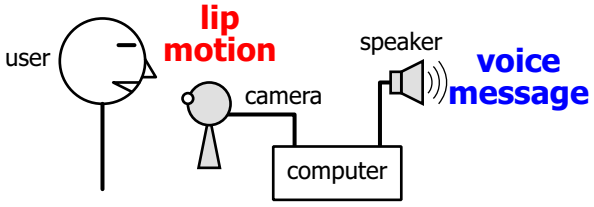



Figure 1: Overview of our transducer system.

which recognizes the mouth shape when the user input one of the key. They used the headset typed camera and applied the thresholding method to detect mouth region. Saitoh et al. proposed the character input system [3]. Their system is based on automatic oral shape recognition detected the stationary oral shape motion unlike Lyons' method. The user can input the character using oral shape and simple head motion. Kato et al. proposed the communication support system [4] similar to [3]. Their system can output one of the voice messages by choosing from the hierarchical 75 conversation voice messages using oral shape recognition. Since [3, 4] are based on five vowel recognition, the identifiable pattern is few. Hence, their operation becomes complicated and it is difficult to development to a high functional application.

## 2 Overview of proposed system

There are two types, a contact type and a non-contact type, to acquire the mouth region. In the case of former type, e.g. a headset type camera which located in front of the mouth, the system can always get stable image without being affected by the movement of the face of the user. However, it is necessary to attach it and gives the user a burden. Therefore, in this research, we employ the latter type. We use a desk camera. The overview of our system is shown in figure 1. The user inputs the system to the lip motion and the system outputs the voice message based on the recognized result.

It is desirable to obtain the large lip region in acquired image to measure the accurate feature. However, there is a problem of the non-contact type which affected easily by the facial movement of the user. Moreover, there is another problem that lip region is over the image size by large opened mouth. Thus, we apply the face image showing in figure 2 for coping with the face and mouth movement and measurement of feature precisely.

(a) closed-shape

(b) a-shape

Figure 2: Acquired image.

(a) closed-shape

(b) a-shape

Figure 3: Extracted results by AAM.

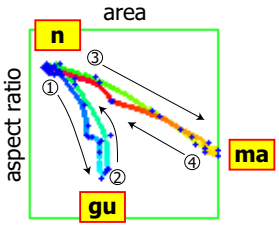



Figure 4: Trajectory feature of "Gunma".

# 3 Lip region extraction

## 3.1 Detect reference measurement

Despite the user utters to be sat down in a straight posture, the position, the size, and inclination of the lip in the image might be changed according to the facial movement when uttering. It is need to prevent varying the feature. To solve this problem, we set the reference measurement. The lip region is not suitable for the reference by changing the shape when uttering. On the other hand, the nostrils do not have a big change during uttering. Then, we set two nostril regions located in the upper side of the lip as the reference measurement.

To detect nostril, we apply the circular separability [5]. We apply this method in the observation area, and a place which filter's output is the biggest detected, as the nostril in both left and right side. Here, when the nostril is detected in the previous frame, the observation area is located based on detected result, otherwise, this area is located the central upper side of the image.

We denote the distance between two detected nostrils $d_n$, and an angle between two nostril $\theta_n$. These values are used for the normalization process described in the next section.

## 3.2 Normalization

We correct the change of the feature by changing the distance $UC$ between the user and camera with $d_n$, and by changing the rotation of the face with $\theta_n$. Based on $d_n$ and the standard distance $d_n^*$, we calculate the scale ratio $s = d_n^*/d_n$, then we apply affine transformation with $s$ and $\theta_n$.

## 3.3 Active Appearance Model [6]

Active Appearance Model (AAM) was proposed by Cootes et al. This method is an iterative fitting algorithm, and it can only deform to fit the target object in ways consistent with the training set. This method extracts two or more object regions simultaneously, and contains a statistical model of the shape and gray level appearance of the object.

## 3.4 Region extraction by AAM

To extract lip region, our AAM model consisted of 40 control points as shown in figure 3. 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. In this paper, we define the inside region of the external lip contour as the external lip region, and the inside of the internal lip contour as the internal lip region.

# 4 Features and recognition

## 4.1 Detect utterance frames

This paper is focused on the word recognition. It is natural to detect the closed lip shape (called closed-shape) that can observe before and after utterance. But, the closed-shape is appeared when the user utter a syllabic nasal or a bilabial. To distinguish between the closed-shape during utterance and closed-shape before and after utterance, we set the threshold time $t_c$. That is, the appearance time of the closed-shape is longer than $t_c$, the system is considered he closed his lip before and after the utterance.

Concretely, $\boldsymbol{P}_t$ is denoted the measured feature vector at time $t$, and $\boldsymbol{P}_c$ is denoted the feature vector of the closed-shape. The condition $(\left|\boldsymbol{P}_t - \boldsymbol{P}_{t-f}\right| \leq T) \cap (\left|\boldsymbol{P}_t - \boldsymbol{P}_c\right| \leq T)$, $(f = 1, 2, \cdots, F)$ is satisfied, the system is considered as the closed-shape without utterance. Where $T$ is threshold value in the feature space, and $F$ is the number of frame in $t_c$[s].

## 4.2 Feature

To recognize the lip motion, we calculate the Trajectory Feature (TF) proposed by Saitoh et al [7]. TF is a motion feature which the feature vector of each frame is plotted into the feature space, the plotted point is interpolated by B-Spline curve, and the motion is expressed a time change of the trajectory. Figure 4 shows a sample TF of Japanese word "Gunma". In this figure, the horizontal axis expresses the area of the internal lip region, and vertical axis expresses the aspect ratio of the internal lip region. The blue point is the feature vector of each frame, and rainbow curve is interpolated result by B-Spline.

## 4.3 Recognition method

As the recognition method of TF, we apply DP-matching similar to the reference [7]. With two TFs, $X$ which is the target unknown word TF and $R_w$ which is registered known word TF, the distance $D(X, R_w)$ is computed. We obtain the minimum distance $D(X, R_w)$, that means, the resulting word is the word $w$ which belongs to $R_w$.

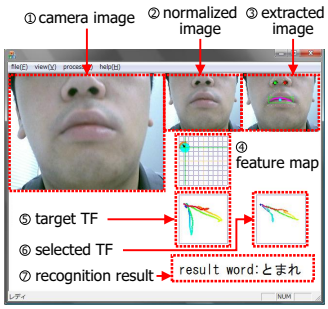



Figure 5: Main window of prototype system.

# 5 Experiments

## 5.1 Prototype system

This paper developed the prototype system which recognized the inputted word by user utterance and output a voice message. Figure 5 shows the main window of the system. In this window, ① is the acquired image, ② is the normalized image described in **3.2**, ③ is the extracted image described in **3.4**, ④ is the time change of the feature vector described in **4.2**, ⑤ is the target unknown TF described in **4.3**, ⑥ is the most similar TF, ⑦ is the recognition result.

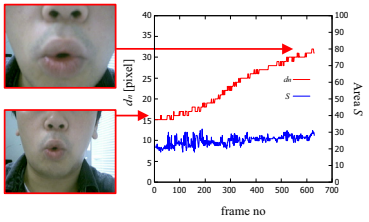
In our system, we use a USB camera (Logicool, Qcam Pro 3000) and the size of the image is $160 \times 120$ pixel. The averaging processing time per one frame was 29ms with a DOS/V PC (CPU: Core2 Duo 2.33GHz). In addition, we set the standard $d_n^* = 20$ pixel for normalization. For the voice synthesis, we use the AquesTalk [8] made by Aquest Corp.

## 5.2 Robustness evaluation for face movement

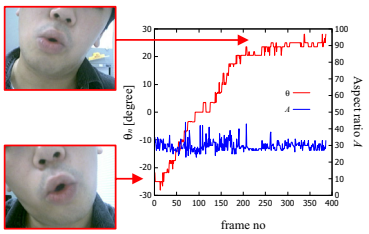
In the traditional research concerned lip reading, the number of researches used the stable utterance scene taken in the specific environment. If it is short utterance like the word, it is easy to give instruction to the person to utter it so as not to move the face for several seconds. However, because this research supposes real-time process, the system acquires the non-utterance scene. The total time of the experiment will be over several minutes. In this case, it is a burden for the person to make stilled state without changing a face in a long time. Therefore, two kinds of movement, (1) the movement to change the distance $UC$, (2) the rotational movement for the optical axis of the camera of the face, were evaluated.

(1) Movement to change the distance $UC$: In this experiment, the user changes $UC$ little by little continuously from 25cm to 12cm with the mouth shape kept constant. This means that the distance $d_n$ between the nostril changes from 15 pixel to 32 pixel in the image space. The target mouth shapes were six typical shapes these are five Japanese vowel shape and closed-shape. We analyzed the area $S$ which used for TF.

Among six shapes, the experimental result of the u-shape is shown in figure 6(a). The horizontal axis is the number of frame, the left vertical axis is $d_n$, and right vertical axis is $S$. In this figure, the red thin line is expressed $d_n$, and blue thick line is expressed




(a) Movement to change the distance $UC$.




(b) Rotational movement.

Figure 6: Experimental results for face movement.

$S$. From this figure, even if $UC$ changes, it is found that $S$ is approximately constant. As the evaluation of quantitative value, the average of $S$ in six shapes was 34.0 pixel and its standard deviation was 4.7 pixel.

(2) Rotational movement: In this experiment, the user rotated his face little by little continuously from $-30°$ to $30°$ with having kept a constant mouth shape like a previous experiment. It was obvious that $S$ did not change for this movement and we analyzed the aspect ratio $A$.

The experimental result of the u-shape is shown in figure 6(b). The horizontal axis is the number of frame, the left vertical axis is $\theta$, and right vertical axis is $A$. In this figure, the red thin line is expressed $\theta_n$, and blue thick line is expressed $A$. From this figure, even if $\theta_n$ changes, it is found that $A$ is approximately constant. As the evaluation of quantitative value, the average of $A$ in six shapes was 24.2 and its standard deviation was 8.9.

From above experimental results, it was verified that the normalization process is robust for the facial movement. But, as for the other movements, such as, the face up and down, and turn to the left and right, the appearance of the mouth is different compared with the frontal mouth shape. Therefore, these problems remain as future work.

## 5.3 Word recognition

In this experiment, we set two word groups, one is the wheelchair control fourteen words ("susume", "sagare", "migi", "hidari", "migikaiten", "hidarikaiten", "tomare", "sukoshisusume", "sukoshisagare", "sukoshimigi", "sukoshihidari", "hayaku", "osoku", and "mouikkai") (called group I), the other is the 47 prefecture name in Japan (called group II). Three physically unimpaired persons (A, B, and C) were cooperated in this experiment. For all subjects, fifteen samples per each word were recorded as the learning data, respectively. In the recognition experiment, the subject uttered ten times of each word, and we computed the average recognition rate. Here, the

learning data and experimental data of AAM and TF come from identical subject. That is, this experiment is speaker dependent experiment. As for the other parameters, we set $T = 10$ pixel and $t_c = 1.5$ sec. based on the experimental result of **5.2**. This value corresponds to $F = 50$ in the same computer. In addition, $UC$ is almost 20cm in this experiment.

The average recognition rates of A, B, and C of group I were 95.7%, 90.7%, and 83.6%, respectively. To analysis the recognition result in detail, we computed the Confusion Matrix (CM) which expressed a tendency for wrong recognition to cause. As the result, ten words obtained more than 90% recognition rate. Oppositely, the word "sukoshisagare" was obtained the lowest recognition rate of 57%, and 27% was misrecognized to the word "sukoshihidari".

The average recognition rates of A, B, and C of group II were 89.6%, 79.8%, and 78.5%, respectively. Eight words, such as "Hokkaido", "Aomori", were obtained the recognition rate of 100%. Oppositely, the word "Saga" was obtained the lowest recognition rate of 26.7%. This word was misrecognized to the word "Nara", "Niigata", and "Akita". In other words, the word "Nara" obtained 30.0%, four words, such as, "Hyogo", "Kyoto", obtained 60.0%.

### 5.4 Process time

In the proposed system, the averaging processing time is 29ms described in **5.1**. This time is from image capturing to region extraction. We also measured the utterance time $t_u$ and recognition time $t_r$ which is the time from the end of the user utterance to the end of the recognition process. These two times of each subject are shown in table 1(a). In this table, PC1 is a DOS/V PC (CPU Intel Core2 Duo 2.33GHz), and PC2 is a DOS/V PC (CPU: AMD Athlon 64x2 Dual Core Processor 2.01GHz).

$t_u$ is depended on the utterance speed of the subject, and this speed of C was slower than A and B. On the other hand, $t_r$ is depended on the computer. Because the number of words is difference between group I and II, the time difference was occurred. Though, A and B used the same PC, the difference is occurred in $t_r$, and A was 2.1 times of B in group II in particular. We investigated this problem and found that this was related to the length $L$ of TF. The length of TF is shown in table 1(b). $L$ of A is 1.4 times of B. This means, both the target unknown word TF and the known word TF of the learning data is 1.4 times large, and the computational cost of the DP-matching is $1.4 \times 1.4 = 1.96$. Thus, the time difference was occurred.

### 5.5 Discussion

In the previous experiment **5.3**, all subjects sat on the chair and experimented by natural posture. There was no big movement at short utterance time because it was word recognition. Then, we carried out additional experiment the word recognition experiment to utter while intentionally moving the face to subject A. As the result, the average recognition rate of group I was 81.4%.

Table 1: Process time and length of TF.

(a) Process time.

| subject | Computer | group I | | group II | |
|---|---|---|---|---|---|
| | | $t_u$[s] | $t_r$[s] | $t_u$[s] | $t_r$[s] |
| A | PC1 | 2.55 | 0.37 | 3.41 | 2.78 |
| B | PC1 | 3.78 | 0.32 | 3.33 | 1.31 |
| C | PC2 | 6.34 | 1.65 | 5.53 | 4.63 |

(b) Length of TF.

| subject | Computer | group I | group II |
|---|---|---|---|
| A | PC1 | 235.5 | 361.7 |
| B | PC1 | 233.1 | 257.9 |
| C | PC2 | 481.3 | 406.7 |

## 6 Conclusion

This paper developed the speaker dependent novel transducer which recognize the lip motion of the word utterance and output the recognized word as the voice message. Though, the base of our system is word lip reading method, our system was adopted the facial movement of the user which does not evade in the real time process, and was verified that our method is robust for it. With three subjects, we obtained the average recognition rate of 90.0% and 82.6%, for fourteen words group and 47 words group, respectively.

In our experiment, we cooperated with three subjects of the physically unimpaired person. The actual user is assumed the speech handicapped person and our final aim is to development the communication support system for these person. Therefore, we will challenge to carry out with actual user. Moreover, our prototype system has some restriction for the facial movement, such as, the face up and down. In the future, we will reduce the restriction for the user.

## Acknowledgement

This research was partially supported by JSPS Grant-in-Aid for Scientific Research (C) 19500476.

## References

[1] I. Matthews, et al.: "Extraction of Visual Features for Lipreading," *IEEE trans. on PAMI*, vol.24, no.2, pp.198–213, 2002.

[2] M. J. Lyons, et al.: "Mouthtype: Text entry by hand and mouth," *Proc. of Conference on Human Factors in Computing Systems (CHI2004)*, pp.1383–1386, 2004.

[3] T. Saitoh, et al.: "Characters entry system based on oral pattern shape," *IEICE Technical Report*, vol.107, no.427, pp.23–28, 2008. (written in Japanese)

[4] T. Kato, et al.: "Communication system based on oral shape recognition," *IEICE Technical Report*, vol.107, no.433, pp.99–104, 2008. (written in Japanese)

[5] O. Yamaguchi and K. Fukui: "Smartface – A robust face recognition system under varying facial pose and expression," *IEICE trans. on Inf. and Syst.*, vol.E86-D, no.1, pp.37–44, 2003.

[6] T. F. Cootes et al.: "Active appearance models," *Proc. of European Conference on Computer Vision (ECCV1998)*, no.2, pp.484–498, 1998.

[7] T. Saitoh, et al.: "Analysis of efficient lip reading method for various languages," *Proc. of International Conference on Pattern Recognition (ICPR2008)*, 2008.

[8] http://www.a-quest.com/aquestalk/index.html.