Automatic Lip Reading for Inability-to-Talk Patient during Mechanical Ventilation

Yudai Nagano, Ryuhei Sakurai, Yu Kawazoe, Kyohei Miyamoto, Hirotake Yamazoe, and Joo-Ho Lee

***Abstract*—In this paper, we propose a lip reading system for patients who have inability to talk during mechanical ventilation. Existing automatic lip reading system cannot be used for ICU patient, because their endotracheal tube causes visual occlusion. We are aiming to read patient's lip by using mouth model that is attached endotracheal tube for this problem. In this paper, we defined two models of mouth shape. One model represents mouth of ordinary face with a set of keypoints extracted by facial keypoints detector. The other model represents mouth of ICU patient. In addition, we compared two models in terms of sequential labeling of visemes and isolated word recognition.**

***Index Terms*—Automatic lip reading, visual speech recognition, hidden Markov model, viseme.**

I. Introduction

In this paper, we propose a verbal communication support system for patients who have inability to talk during mechanical ventilation. For the patient, it is important to communicate by using language to tell their intention. Patients who suddenly have to be in ICU by accidents also need to tell their health condition to medical staff. However, most of ICU patients cannot speak since they are artificially ventilated to keep their lives and their vocal cords are restricted. Although, in the past, a ventilated patient was kept under deep sedation, lighter sedation is common nowadays in order to shorten recovery time from the ventilator support and have a good prognosis. Because of this, it has rapidly increased that communication opportunities between medical staff and ICU patients. In such a communication, word-cards or writings are often used. However, these methods usually cannot be used because most of ICU patients are in serious conditions, to which they can move only facial muscles or some fingertips due. Therefore, actually, medical staff has to gaze patient's mouth and do lip reading, which is difficult to get what the patient is saying, and it takes much time. This difficulty costs mentally and physically to both the patients and the medical staff. Thus, communication support system has been strongly desired.

In this paper, we propose a communication support system by using automatic lip reading. Automatic lip reading is the technique of speech recognition by using only visual information. By using the proposed automatic lip reading system, it is expected that patients can verbally communicate with their medical staff without special training. In general, visemes are used as minimum unit of estimation [1], and automatic lip reading systems first extract some features from the mouth in a video of speech and feed the features to a speech recognizer [2]. However, in the case of ICU patients, it is difficult to extract features from their mouth (Fig. 1) because the patient is intubated a endotracheal tube into his windpipe and the tube is fixed to his mouth with tape which causes visual occlusion. Furthermore, the retainer disturbs the movement of the mouth. We are firstly aiming to read patient's lip by using mouth model that is attached endotracheal tube for this problem. In this paper, we define two models of mouth shape. One model represents mouth of ordinary face with a set of keypoints extracted by facial keypoints detector. The other model represents mouth of ICU patient, and it is assumed that one side of the mouth is occluded by the retainer. In addition, we compare two models in terms of sequential labeling of visemes and isolated word recognition to elucidate the influence of lip-reading accuracy of lack of mouth model.

Fig. 1. Simulated ICU patient.

II. Related Works

An automatic lip reading system mainly consists of two parts which are the training part and the estimation part. In training part, it extracts features from a lot of videos of speech and builds the estimator of viseme sequence. For the estimator, Hidden Markov Model (HMM) is often used. HMM is commonly used in the field of speech recognition. In estimation part, the features are extracted from input data, and these features are input into the viseme estimator. For example, there is a method that extracts the mouth shape by using Sampled Active Contour Model and uses its shape value as features. This method can recognize vowels with high accuracy [3]. However, the method has a bottleneck of the accuracy of shape detector.

III. Proposed System

In real ICU, the medical staff ask to patients, for example,

Manuscript received September 21, 2015; revised May 9, 2016.

Yudai Nagano, Ryuhei Sakurai, Hirotake Yamazoe, and Joo-Ho Lee are with Ritsumeikan University, Japan (e-mail: is0081hx@ed.ritsumei.ac.jp, rsakurai@fc.ritsumei.ac.jp, yamazoe@fc.ritsumei.ac.jp).

Yu Kawazoe is with Tohoku University, Japan (e-mail: ukz411@gmail.com).

Kyohei Miyamoto is with Wakayama Medical University, Japan (e-mail: go.go.kyohei.miyamoto@gmail.com).

"what parts of your body are painful?". Such as this question, it is limited to reply for patients. However in case of free talk, it is difficult to expect what patient will tell next. Thus we define the conversation scene between medical staff and ICU patients as Closed-Question (CQ) and Free-Talk (FT). CQ is the scene of asking some questions from medical staff to the patient. FT is the scene of free talk with medical staff and the patient.

We propose the system behave to fit each scene. In particular, in the case of CQ, the proposed system shows a list of most likely words to medical staff when the patient reply to the question from medical staff (Fig. 2). The list may help medical staff to recognize the patient's word which he is trying to tell. Next, in case of FT, the proposed system outputs to medical staff the viseme information that is recognized from the patient's speech without voice (Fig. 3). It may help medical staff to estimate the patient's intention from viseme information.

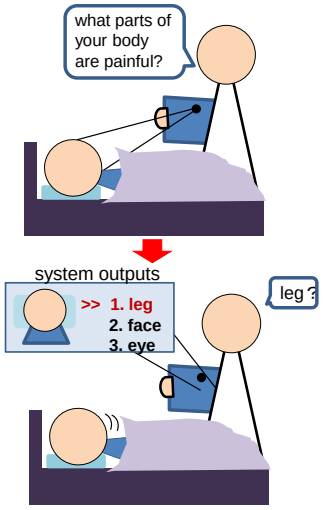


Fig. 2. Closed question case.

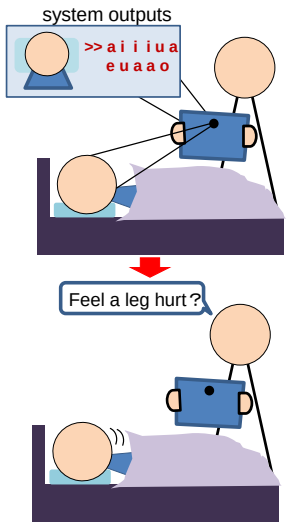


Fig. 3. Free talk case.

IV. System Flow

In this section, we describe the system flow at both CQ case and FT-case. Fig. 4 shows the system outline in CQ case. In this case, we defined viseme as minimum unit of estimation. We referred to Fukuda et al. [4] to define the visemes and add viseme /N/. Table I shows the correspondence between phoneme and viseme. In this case, for the training part, to make viseme estimator we trained each viseme HMM by lots of speech videos. The estimation part extracts features from input data and feeds the features into viseme estimator. The viseme estimator outputs the estimated result as a sequence of visemes. Fig. 5 shows the system outline in FT-case. In FT-case of training part, to make word classifier, we trained each word HMM by lots of videos of isolated words. The estimation part extracts features from input data and feeds into word classifier. The word classifier outputs the result as a word.

TABLE I: Correspondence Table of Phoneme and Viseme

Japanese Phoneme	Japanese Viseme	Japanese Phoneme	Japanese Viseme	Japanese Phoneme	Japanese Viseme
a	a	j	sy	t	t
a:		my		d	
i	i	ky		n	
i:		by		ts	s
u	u	gy		z	
u:		ny		s	
e	e	hy		y	y
e:		ry		k	vf
o	o	py		g	
o:		ch		h	
p	p	dy		N	None
b		sh		q	None
m		w	w		
r	r	f			

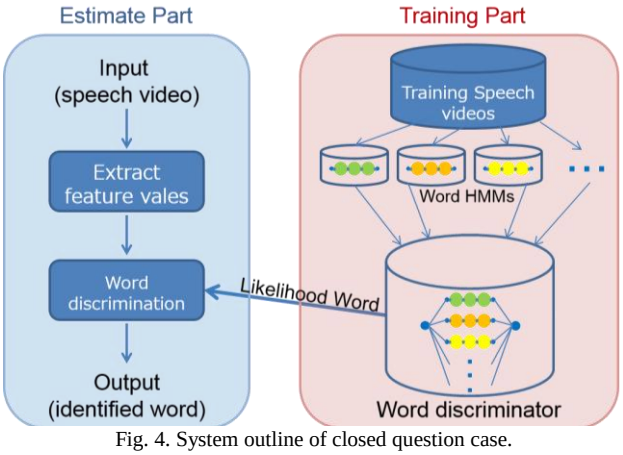


Fig. 4. System outline of closed question case.

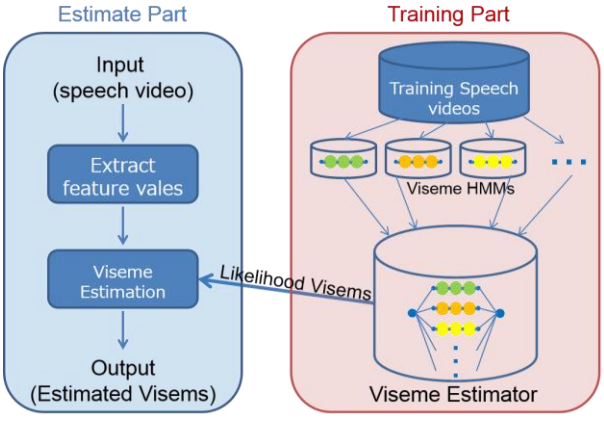


Fig. 5. System outline of free talk case.

A. Speech Video for Training

In this section, we describe a dataset of speech video used to train the viseme estimator and word classifier. The speaker

spoke naturally 5 times of a sets of 492 words [5] in consideration of balance of phoneme and we captured each speech one-by-one. Since durations of consonants are very short time, we have to use high frame rate camera in order to capture the consonants. Thus, we used Gopro-Hero3 camera. We set the video resolution 1280×780, and 120 FPS. The speaker sits in front of the camera, and the distance between the speaker and the camera is about 50cm. We set the camera to capture the whole face of the speaker. Fig. 6 (a) shows a moment of speaking phoneme /a/, Fig. 6 (b) shows a moment of speaking phoneme /i/, Fig. 6 (c) shows a moment of speaking phoneme /e/, Fig. 6 (d) shows a moment of speaking phoneme /u/. In addition, we annotated meta information such as viseme segment to speech video. In general, although the labeling is done by expert while watching movements of speaker's mouth, this method takes much cost. Thus we used the Japanese phoneme segmentation kit [6] to the voice that extracted speech video and used its outputs as viseme segment information (Fig. 7).

(a) Phoneme/a/ (b) Phoneme/i/

(c) Phoneme/e/ (d) Phoneme/u/

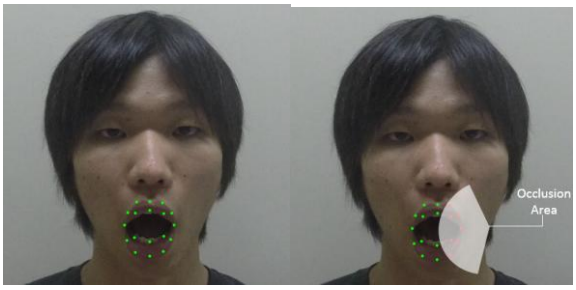
Fig. 6. The speaker.

B. Feature Extraction

We used the facial feature points which can be obtained by using incremental iPar-CLR metod [7] and used relative coordinate of feature points around mouth from the point of the middle of eyeblows as features. Fig. 8 (a) shows the using mouth points and we defined this as full mouth model. It consists of 18 points. On the other hand, we need to consider the case of ICU patients. As mentioned above, ICU patients have endotracheal tube and supported tape at their mouth. Thus we defined the mouth model of ICU patients as occluded mouth model (Fig. 8(b)). It consists of the 11 points among the points of full mouth model, considering the occluded area. The full mouth model's feature dimension for each picture is 36 and occluded mouth model's feature dimension is 22.

C. Training Viseme and Word HMM

In this section, we describe the training HMM to make viseme estimator and word discriminator. We used left to right HMM with mixture of Gaussians as emission model. We used Hidden Markov Tool Kit [8] for making HMM.



(a) Full mouth model (b) Occluded mouth model

Fig. 8. Defined mouth model.

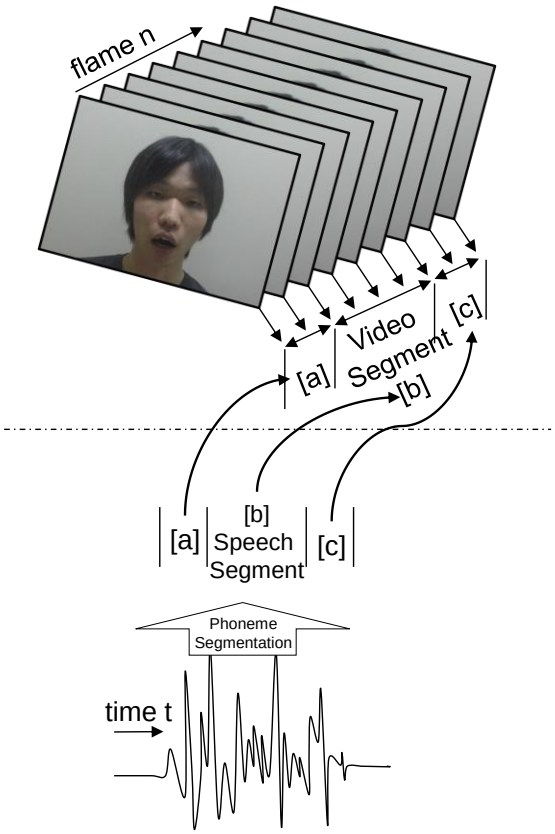


Fig. 7. Labelling to speech video.

V. Experimentation

In this section, we describe the experiments for viseme estimator and word classifier. In addition, we compared the features extracted from both full mouth model and occluded mouth model.

A. Viseme Estimation

We used cross validation method to evaluate viseme estimator by using the speech video data that described in section 4.1. The number of speaker is one. When 4 sets was used for training, and then the rest 1 set was used for estimation. We repeated the procedure 5 times and evaluate the viseme estimator from the view point that is the correct rate of estimated labels.

Table II shows the correct of viseme estimator with vowels and consonants by each mouth model. We calculated the correct rate that total number of correct located label that estimated per total number of correct label, with each viseme. We found that in case of vowels estimation with full mouth model correct mostly. Similarly, in case of vowels estimation with occluded mouth model correct mostly. It is assumed that

from this evaluation results the symmetry of movement of mouth is kept when we speak. For this reason, we found that occluded mouth model, which is expected to be extracted by ICU patients, can be used to estimate vowels.

Table III and Table IV show the confusion matrices which contain all visemes, insertion and deletion, by using each mouth model. The each column means the true label of viseme and the each row means the estimated label. The bottom row (insert) means the number of insertion error for each viseme and the rightmost column (deleat) means the number of deleation error of each viseme. At this result, it is assumed that viseme /a/ and /e/ is difficult to be distinguished, because of their visual similarity. Similarly, it can be said for viseme /vf/ and /sy/.

TABLE II: VISEME CORRECT RATE

	Full Mouth Model (%)	Occluded Mouth Model (%)
Vovels	76	73
Consonants	49	51
Total	59	59

TABLE III: CONFUSION MATRIX OF VISEME UNDER FULL MOUTH MODEL

	a	i	u	e	o	p	r	sy	w	t	s	y	vf	N	Delete	Correct(%)
a	255	11	0	37	2	2	4	1	0	3	2	1	4	1	79	77.62
i	12	172	2	8	1	3	5	6	1	6	4	1	5	7	92	71.48
u	4	2	198	3	7	7	5	3	5	11	8	1	8	12	109	71.42
e	31	8	1	153	0	4	3	2	1	1	1	0	2	1	67	71.02
o	1	1	4	1	302	3	3	2	5	2	3	1	5	3	57	90.36
p	1	0	0	1	0	202	1	0	0	0	0	0	1	0	15	78.02
r	3	3	4	3	2	2	25	1	2	6	2	2	4	1	70	42.68
sy	4	17	4	4	4	22	5	143	2	13	9	4	11	0	89	58.26
w	1	0	1	1	3	4	0	0	26	1	0	0	1	1	14	63.54
t	5	6	7	6	3	3	8	5	3	50	3	2	7	2	107	42.82
s	2	6	3	2	1	2	3	10	0	5	67	0	5	2	54	58.22
y	1	3	1	1	1	0	2	4	0	1	1	4	2	0	16	19.04
vf	8	9	2	6	4	4	11	8	3	12	9	2	52	2	157	32.32
N	7	4	3	4	3	8	3	1	2	3	2	1	5	38	97	44.82
Insert	13	9	15	12	16	36	22	10	14	22	9	6	24	32		

TABLE 4: CONFUSION MATRIX OF VISEME UNDER OCCLUDED MOUSE MODEL

	a	i	u	e	o	p	r	sy	w	t	s	y	vf	N	Delete	Correct(%)
a	239	9	1	34	1	4	5	1	1	4	4	1	5	4	91	75.98
i	5	167	2	6	1	3	5	4	1	8	6	3	4	13	95	71.76
u	3	2	177	4	6	12	7	2	10	7	9	8	7	8	121	66.40
e	26	11	1	130	1	4	7	1	1	3	1	2	1	3	83	65.84
o	1	1	6	1	284	3	4	2	9	3	4	4	4	2	64	86.68
p	0	0	0	0	0	203	1	0	1	0	1	0	0	0	14	78.14
r	3	2	0	2	2	1	34	1	3	3	4	4	3	1	67	53.40
sy	2	15	1	2	4	27	11	121	3	10	17	14	9	3	93	50.26
w	0	0	0	0	4	2	1	0	33	1	0	0	0	0	11	76.32
t	3	6	5	3	2	4	14	3	6	43	5	5	5	2	112	39.72
s	2	7	3	0	0	1	4	9	0	4	77	3	2	2	48	66.54
y	0	3	0	0	1	0	3	3	0	1	1	7	1	0	18	30.38
vf	5	7	2	4	2	2	19	8	4	10	10	9	36	5	164	26.10
N	9	5	2	3	2	10	5	1	4	4	2	3	3	36	92	39.94
Insert	13	7	7	9	7	57	37	6	25	21	11	15	17	38		

B. *Word Discrimination*

We evaluated the word classifier. First, we picked up 30 words at random from the dataset which is balanced in terms of phonemes. Then we made word classifier by training with 5

examples of speech with each word HMM. We evaluated the word classifier to input new 5 examples to this word classifier. All word HMMs are constructed with same number of states and mixture components as hyperparameters. We found the best hyperparameter by grid search from the viewpoint of the highest accuracy of word classifier. Fig. 9 shows the correspondence between number of words and word accuracy rate. We calculated the accuracy rate that total number of estimated word that corrected per total number of trials, with each word. From the result, we found that the increase of the number of words tend to decrease the accuracy rate of words. In addition, we found the similar results between full mouth model and occluded mouth model. Thereby, this results signify that we can get the similar accuracy for word by using occluded mouth model from real ICU patients in automatic lip reading system.

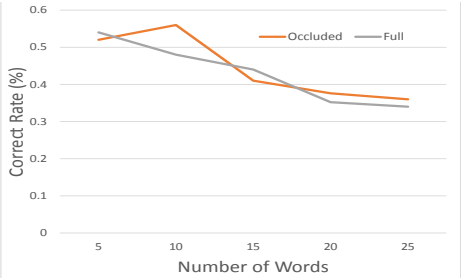


Fig. 9. Word accuracy rate.

VI. Conclusion

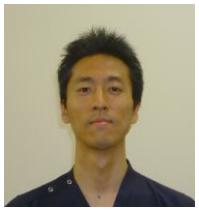
In this paper, we proposed the communication support system based on automatic lip reading. We compared the mouth model extracted from healthy person and the mouth model which is expected to be extracted from ICU patients. In each evaluation about viseme estimation and word classification, we found the similar results from each mouth model. Thereby, this results signify that we can use occluded mouth model from real ICU patients in lip reading system. From the result of evaluation of viseme estimator, we have to make more discriminative feature representation due to identify viseme /a/ and /e/. In this paper, we made speech video for training with only one person. Thus we have to collect more speech data from than one person and adapt the estimator/classifier for general person. In addition, since ICU patients who have endotracheal tube cannot move their mouth freely, we have to consider the movement of mouth with real patients. We then collect speech video that spoken by simulated ICU patients that is attached endotracheal tube, in order to fit the acoustic models for the real patients. After that, we intend to train acoustic models by using that speech videos.

References

[1] C. Luca and H. Naomi, "Phoneme-to-viseme mapping for visual speech recognition," in *Proc. ICPRAM*, vol. 2, pp. 322–329, 2012.
[2] Phoneme segmentation kit. [Online]. Available: http://julius.osdn.jp/
[3] T. Saitoh, M. Hisagi, and R. Konishi, "Analysis of features for efficient japanese vowel recognition," *IEICE TRANS.INF and Systems*, vol. 11, pp. 1889–1891, 2007.
[4] F. Yumiko and H. Shizuo, "Characteristic of the mouth shape in the production of japanese-stroboscopic observation," *IEICE*, pp. 259–265, 1978.
[5] K. Tanaka, S. Hayamizu, and K. Ohta, "The etl speech database for speech analysis and recognition research," *ICSLP-90*, vol. 24, no. 7, pp. 1101–1104, 1990.
[6] Phoneme segmentation kit. [Online]. Available: http://julius.osdn.jp/
[7] Asthana and Akshay, "Incremental face alignment in the wild," *Computer Vision and Pattern Recognition (CVPR)*, pp. 1859–1866, 2014.
[8] Htk. [Online]. Available: http://htk.eng.cam.ac.uk/

Yudai Nagano received the B.E degrees in information science and engineering, Ritsumeikan University in 2013. He is now a master course student at the Graduate School of Information Science and Engineering, Ritsumeikan University. His research interests include visual speech recognition and computer vision.

Ryuhei Sakurai received the B.E. and M.E. degrees in engineering from Ritsumeikan University in 2005 and 2008, respectively. He then finished his Ph.D. program without dissertation, Graduate School of Science and Engineering, Ritsumeikan University in 2012. He is now an assistant with the College of Information Science and Engineering, Ritsumeikan University. His research interests include computer vision and machine learning.

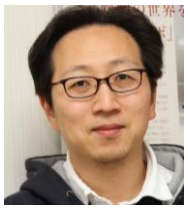


Yu Kawazoe received a M.D. degree form Kyoto Prefecture University of Medicine in 2003 and a Ph.D. degree in medicine from Wakayama Medical University in 2015. He was an assistant professor with the division of emergency and critical care medicine, Wakayama Medical University during the years from 2005 to 2015. He is now an assistant professor with the division of emergency and critical care medicine at Tohoku University.

His research interests include monitoring systems and treatment in Critical Care Medicine. He is a member of the Japanese Association for Acute Medicine, the Japanese Society for Emergency Medicine, the Japanese Society of Intensive Care Medicine, the Japan Surgical Society and the Japanese Association for the Surgery of Trauma.

Kyohei Miyamoto received a M.D. degree from Jichi Medical School in 2006. He is an assistant professor with the division of emergency and critical care medicine, Wakayama Medical University from 2015. His research interests include infectious disease and nutrition in critical care medicine. He is a member of the Japanese Society of Intensive Care Medicine.

Hirotake Yamazoe received the B.E., M.E., and Ph.D. degrees in engineering from Osaka University in 2000, 2002, and 2005, respectively. He was with the Advanced Telecommunications Research Institute International (ATR) during 2005–2011, the Institute of Scientific and Industrial Research, Osaka University, during 2011–2012, and Osaka School of International Public Policy, Osaka University, during 2012–2015. He is now a lecturer with the College of Information Science and Engineering, Ritsumeikan University. His research interests include computer vision and wearable computing. He is a member of the IEICE, IPSJ, ITE, HIS, VRSJ, ACM, and IEEE.



Joo-Ho Lee received the B.E and M.E degrees from Korea University and the Ph.D. degree from University of Tokyo. He was JSPS researcher in Tokyo University during 2000-2002, and post-doctoral researcher in Tokyo University during 2002-2003, and research associate in Tokyo University of Science during 2003-2004. He is now a professor with the College of Information Science and Engineering, Ritsumeikan University. His research interests include Intelligent Space, Service Robots and Computer Vision. He is a member of JSME, RSJ, IEICE, SICE, IEEJ, and IEEE.