IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 17, 2016, accepted January 4, 2017, date of publication January 9, 2017, date of current version March 6, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2649838

Development of Novel Lip-Reading Recognition Algorithm

BOR-SHING LIN[1], (Member, IEEE), YU-HSIEN YAO[2], CHING-FENG LIU[3,4], CHING-FENG LIEN[5,6], AND BOR-SHYH LIN[2], (Senior Member, IEEE)

[1]Department of Computer Science and Information Engineering, National Taipei University, New Taipei City 23741, Taiwan
[2]Institute of Imaging and Biomedical Photonics, National Chiao Tung University, Tainan 71150, Taiwan
[3]Department of Medical Research, Chi Mei Medical Center, Tainan 71004, Taiwan
[4]Graduate Institute of Medical Sciences, Chang Jung Christian University, Tainan 71101, Taiwan
[5]Department of Otolaryngology–Head and Neck Surgery, E-Da Hospital, Kaohsiung 82445 Taiwan
[6]I-Shou University, Kaohsiung 84001, Taiwan

Corresponding author: B.-S. Lin (borshyhlin@gmail.com)

This work was supported by the Ministry of Science and Technology of Taiwan, under Grant MOST 105-2221-E-009-013 and Grant MOST 105-2221-E-305-001.

ABSTRACT Total laryngectomy is a common treatment for patients with advanced laryngeal and hypopharyngeal cancer, but it is also a result from the loss of the natural voice and directly affects the basic communication functions in daily life. Reconstructing the basic communication function is an important issue for these patients after total laryngectomy surgery. Recently, the image processing technique for lip-reading recognition has been widely developed and applied in various kinds of applications. It is also one of the possibly alternative approaches to reconstructing the basic communication function for these patients after total laryngectomy surgery. Although many human lip-reading recognition methods have been developed to detect lip contour precisely, detecting pronouncing lip contour effectively is still a difficult challenge. In this paper, a novel lip-reading recognition algorithm was proposed to recognize English vowels from the lip contour when speaking. Here, several criteria for detecting the mouth region of interest (ROI) were designed to reduce the error rate of detecting the mouth ROI and lip contour. Moreover, several lip parameters, including the width, height, contour points, area, and the ratio (width/height) of lips, were used to recognize the lip contour and English vowels when speaking. The advantages of the proposed method are that it could detect the mouth ROI automatically, reduce the influence of individual differences, such as the individual lip shape or makeup effect, and it also could perform a good performance without pretraining. Finally, the performance of lip-reading recognition under different backgrounds and individual differences was also tested, and the accuracy of the proposed algorithm on lip-reading recognition was over 80%.

INDEX TERMS Laryngectomy, lip-reading recognition, mouth region of interest, visual-only speech recognition, vowels recognition.

I. INTRODUCTION

Laryngeal cancer is a disease that malignant cells form in the larynx tissues, and it caused about 800,000 deaths each year [1]. Total laryngectomy (TL) is a common treatment for these specific patients with advanced laryngeal and hypopharyngeal cancer [2]–[4], but it also results in the loss of the natural voice and directly affects the basic communication functions in daily life [5]. Reconstructing the basic communication function is an important issue for these patients after total laryngectomy surgery. Recently, the image processing technique for human lip recognition has been widely developed and applied in various kinds of applications. It might contain the potential of reconstructing the basic communication function for the patients with total laryngectomy surgery.

Recently, the image processing technique for human lip recognition, which is able to automatically detect and analyze the inconstant shape of the human lip and has the ability to distinguish whether the user is talking or not for real-time, has been widely developed, such as audio-visual speech recognition (AVSR) [7], visual only speech recognition (VSR) [8], [9], speaker identification [10]–[12], intelligent human–computer interaction [13], human expression recognition [14], lip segmentation for VSR or AVSR [15], [16], vision based voice activity detection (VVAD) [17], *et al.* Here, audio-visual speech recognition system analyzed the amplitude and frequency of speech sounds and extracted the lip information from the image sequences simultaneously to performed speech recognition. Visual only speech recognition was designed to detect the opening or closing state of the mouth from the image sequences. Speaker identification was developed to recognize the lip shape of the user, and then applied in identity certification security. Human expression recognition was designed to distinguish real or fake smiles, and then to estimate the emotion of the user. The common processing procedure of the above methods for human lip recognition mainly contained three steps: face detection, region of interest (ROI) localization, and visual feature extraction. However, recognizing human lip precisely from the continuous image sequences is still a difficult challenge due to the variations on the basis of lip shapes and hues.

Several approaches of moving human lip recognition have also been proposed in previous studies. In 2004, Eveno *et al.* used the jumping snake approach, which is a matching method for deformable models [18], to detect human lip contour. However, in this method, the location of the desired contour shape had to be selected manually. In 2006, Cetingul *et al.* utilized the explicit lip motion information associated with hidden Markov model (HMM) to do speaker identification and speech recognition [10]. Here, the pre-training procedure was required, and this also increased the inconvenience of use in daily life applications. In 2012, Liu *et al.* proposed a lip tracking approach for speaker identification, that the darkest sites in the grayscale photo was viewed as the mouth corners, and a relatively large value of divergence between neighboring pixels was recognized as contours [11]. However, it was easily affected by the environmental conditions, such as black background. In 2012, Chin *et al.* proposed a region-based active contour model with local information using watershed segmentation was also proposed for lips contour detection [19]. In this method, a pre-defined set of markers were required to divide an image. It might be difficult to satisfy different locations and sizes of the contour shapes. The above mentioned methods could provide a good ability for detecting individual lip contour or simple speech recognition, but they could not precisely recognize each English vowel when speaking. Moreover, they are also easily affected by the influence of the change of the background condition, initial error and dramatically mutative shape of mouth.

In order to improve the above issue, a novel lip-reading recognition algorithm was developed to real-time recognize each English vowel when speaking. In this algorithm, the image pre-processing and several criteria for detecting the mouth region of interest (ROI) were designed to reduce the influence of the variation of the environmental condition and to avoid the error occurrence results from the detected incomplete contour or the remaining part of the background image. After detecting the mouth ROI, several lip parameters, including the total pixel number of the detected lip contour, and the height, width, area and ratio of the lips, were used as the features of lip-reading recognition. Finally, the correlation between these extracted lip parameters and the lip parameters corresponding to each English vowel would be calculated to precisely recognize each English vowel. By using the above techniques, the proposed method could detect the mouth ROI automatically, reduce the influence of individual differences, and it also could performance a good perform without pre-training.

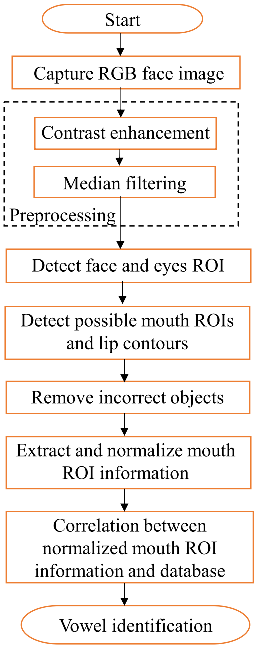

Start
Capture RGB face image
Contrast enhancement
Median filtering
Preprocessing
Detect face and eyes ROI
Detect possible mouth ROIs and lip contours
Remove incorrect objects
Extract and normalize mouth ROI information
Correlation between normalized mouth ROI information and database
Vowel identification

FIGURE 1. Flowchart of proposed lip-reading recognition algorithm.

II. METHODS

A. LIP-READING RECOGNITION ALGORITHM

The flowchart of the proposed lip-reading recognition algorithm was illustrated in Fig. 1. In order to enhance the image contrast for human vision, the image of captured by the webcam would be first processed by extending the original minimum and maximum pixel values to the values of 0 and 255 respectively. Then, the technique of median filtering, which is a nonlinear digital filtering technique and has been widely applied in preserving the contour edges and removing noise, such as salt, pepper noise, speckle noise etc., would be used to reduce noise in the captured image.

Next, both of the regions of the face and the eye pair on each frame image would be automatically detected using a cascade of boosted classifier, proposed by Viola and Jones [15]. The region of the eye pair was only detected when a face was detected on the current image, and it has to be inside the face ROI, as illustrated in Fig. 2.

In order to extract the possible region of the lips, the rationality for the locations of face and eye pair regions have to be checked, and then a sub-region of the mouth [9] could be roughly estimated by

$$(x_2, y_2) = \left(x_0 + \frac{M}{6}, y_0 + \frac{5N}{8}\right) \tag{1}$$

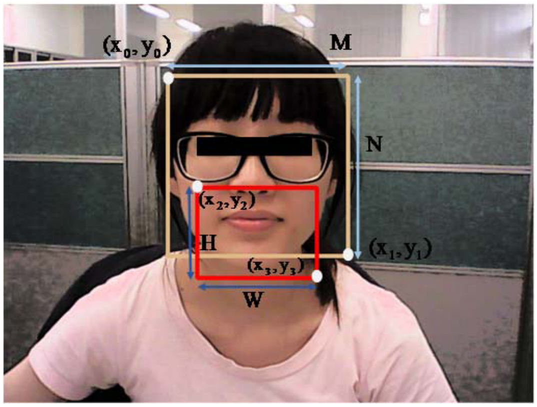


FIGURE 2. Illustration of face and mouth ROI in lip-reading recognition algorithm.

$$(x_3, y_3) = \left(x_2 + \frac{2M}{3}, y_2 + \frac{N}{2}\right) \tag{2}$$

where (x_0,y_0) and (x_1,y_1) denote the origin and the bottom-right positions in the bounding box of the detected face region respectively, and its size is $M \times N$. Here, (x_2,y_2) and (x_3,y_3) denote the origin and the bottom-right positions in the bounding box of the detected mouth, and its size is $W \times H$. In order to more accurately detect the lip edge, these smaller objects which might be not the lip contour, such as moles and freckles, and the relatively larger objects, such as the face contour, would be eliminated by using two thresholds (A_{min} and A_{max}) of the area size. Here, the thresholds A_{min} and A_{max} were defined as 0.01 time and 0.5 time the size of the detected mouth respectively. Some incorrect regions, such as nose, shadow etc., that their size might also be within the reasonable size (A_{min} < region size < A_{max}), also have to be removed.

Besides the check of the reasonable size for the mouth ROI, the horizontal and vertical center locations of the estimated mouth ROI would also be checked. When the horizontal and vertical center locations of the estimated mouth ROI were not within the reasonable range, this region would not be considered as the region of mouth and then would be removed. The reasonable range is between $0.3W$ to $0.7W$ and $0.25H$ to $0.5H$ (within the area of $0.4W \times 0.25H$, as shown in Fig. 3). After checking the region size and its center location simultaneously, the region fitted the checking criterion would be viewed as the mouth ROI.

After detecting the mouth ROI, the information for the lip contour in this region and the bounding box of this region would be used for vowel recognition. The used information included the total pixel number of the detected lip contour, and the height, width, area and ratio of the lip bounding box (the mouth ROI). In order to reduce the influence of the individual difference, the above parameters would be first normalized before vowel recognition. Here, the information of the user's face image would be pre-recorded in the beginning of the system program, and was used as a measure of scale for normalization. Here, the information of the user's face image was obtained from five face images that the user closed his/ her mouth. Next, the correlation between the normalized lip contour information vector and each vowel feature vector would be obtained by Pearson correlation coefficient [16], that can be given by

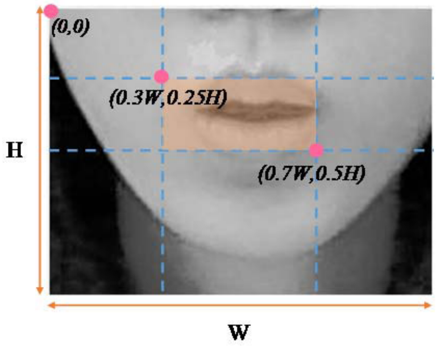


FIGURE 3. Illustration for criterion of reasonably lip center position.

$$\rho_{Pearson}(\mathbf{X}; \mathbf{Y}) = \frac{Cov(\mathbf{X}, \mathbf{Y})}{\sqrt{Var(\mathbf{X})\, Var(\mathbf{Y})}} \tag{3}$$

where **X** and **Y** denotes the current normalized lip contour information vector and one of vowel feature vectors respectively, and the functions of $Cov(.)$ and $Var(.)$ denotes the calculations of covariance and sample standard deviation respectively. The parameters of w, h, a, p and r are used to denote the width, height area, contour points and ratio of the initialized lip contour obtained from the user respectively. Here, the ratio r denotes as w/h. After calculating the correlations corresponding to different vowel feature vectors, the vowel feature vector providing the highest correlation value would be viewed as the output of vowel recognition.

In order to perform lip-reading recognition continuously, the steady state (the steady lip contour) and transient state (the changing lip contour) of speech have to be distinguished. In general, the correlation between the normalized lip contour information vector and one of vowel feature vectors would be larger than 0.5 when the lip contour was under the steady state. Therefore, the output of vowel recognition would be generated when its correlation was larger than 0.5. Otherwise, it would be recognized as the transient state when its correlation was smaller than 0.5. The time-series diagram of lip-reading recognition was illustrated in Fig. 4.

B. IMPLEMENTATION OF LIP-READING RECOGNITION SYSTEM

The proposed lip-reading recognition system mainly consists of a webcam (Logitech C170, Logitech, Taiwan) and a laptop with Pentium 2.94-GHz CPU. The program built in the laptop was developed by using Visual Basic 2012. Here, the webcam was used to capture a series of continuous face images. A real-time monitoring program for lip-reading recognition, developed by visual C# with Open source computer vision library (OpenCV 3.1.0) and Emgu CV 2.3.0, was also built

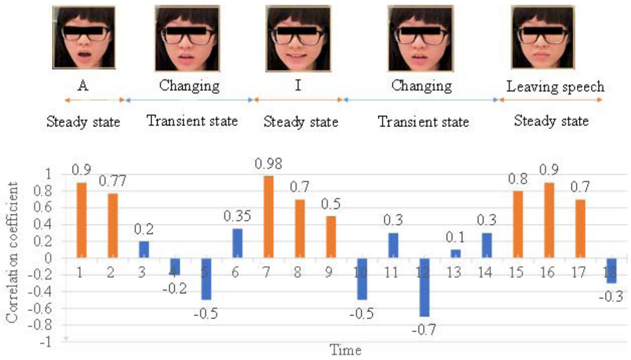
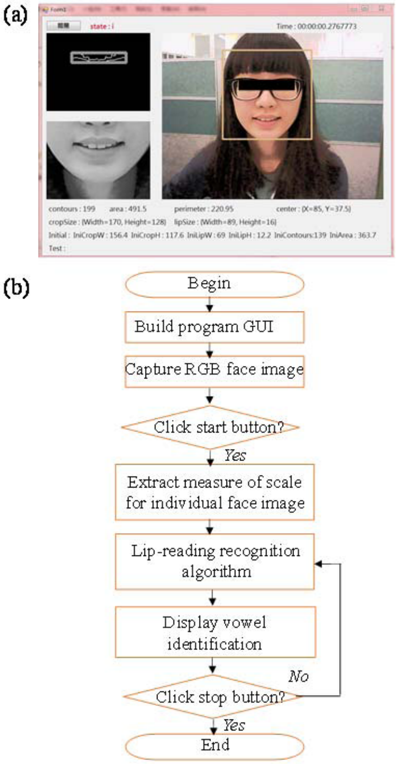


FIGURE 4. Time-series diagram of lip-reading recognition.


(a)
(b)
Begin
Build program GUI
Capture RGB face image
Click start button?
Yes
Extract measure of scale for individual face image
Lip-reading recognition algorithm
Display vowel identification
No
Click stop button?
Yes
End

state : i

Time : 00:00:00.2767773

contours : 199 area : 491.5 perimeter : 220.95 center : {X=85, Y=37.5}

cropSize : {Width=170, Height=128} lipSize : {Width=89, Height=16}

Initial : IniCropW : 156.4 IniCropH : 117.6 IniLipW : 69 IniLipH : 12.2 IniContours:139 IniArea : 363.7

Test :

FIGURE 5. (a) Screenshot and (b) flowchart of real-time monitoring program.

in the laptop. Fig. 5 (a) showed the screenshot of the real-time monitoring program. Here, OpenCV and Emgu CV are used to process image, and Emgu CV is a cross platform .Net wrapper to the OpenCV image processing library, allowing OpenCV functions to be called from .NET compatible languages such as C#, VB, VC++, IronPython etc.

The flowchart of the real-time monitoring program was shown in Fig. 5 (b). In the beginning of this program, the graphic user interface would be first established to allow the user operating this program. Then, the webcam would be enabled to continuously capture the color face image with the sampling rate of 30 frames per second (fps). Before performing the lip-reading recognition algorithm, the program would display ''Close your mouth to initialize!'' to instruct the user to close his/ her mouth, and then after clicking the start button, it would capture five face images continuously to obtained the measure of scale, used for normalizing the lip parameters in the proposed lip-reading recognition algorithm. After extracting the information of the user's face image successfully, this program would display ''Initialization is complete!''. Next, this program would start to perform the lip-reading recognition algorithm continuously to recognize vowel from the captured face image, until clicking the stop button.

TABLE 1. Ground truth of actual and predictive frames in three conditions.

	Vowel 'a'	Vowel 'i'	Vowel 'u'	Vowel 'e'	Vowel 'o'	Silence
Actual frames	248	254	299	891	593	824
Predictive frames	237	242	327	959	556	788

C. EXPERIMENTAL DESIGN

In order to evaluate the performance of the proposed method, several image sequences under different environmental conditions were used for test. There are three types of environmental conditions and four types of individual differences for tests, including one static background (**condition 1**), two dynamic backgrounds (fan rotating background: **condition 2**, and people walking background: **condition 3**), thick-lip makeup effect (**individuality 1**), moustache face (**individuality 2**), and different face angles (30-degree:**individuality 3**, and 60-degree: **individuality 4**). Here, a total of ten participants attended this experiment. Each trial contains a least 10 times vowel utterances. Before evaluating the performance of the proposed method, several parameters of binary classification test have to be first defined, and they were listed as follows: True positive (*TP*) denotes a specific vowel was correctly detected as the specific vowel. False positive (*FP*) denotes other vowels or silence was incorrectly detected as a specific vowel. True negative (*TN*) denotes silence was correctly detected as silence. False negative (*FN*) denotes a specific vowel was incorrectly detected as silence. Here, the parameters of sensitivity (also called true positive rate, TPR), precision (also called positive predictive value, PPV) and accuracy (ACC) were used to evaluate the performance of vowel recognition, and they were defined as follows.

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$PPV = \frac{TP}{TP + FP} \tag{5}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{6}$$

III. RESULTS

Ground truth is a term used in various fields to refer to information provided by direct observation as opposed to information provided by inference. In order to determine the ground truth, each frame was marked after visual inspection,

TABLE 2. True positive rate (TPR), positive predictive value (PPV), and accuracy (ACC) of proposed algorithm on lip-reading recognition under different conditions.

	TP	TN	FP	FN	TPR	PPV	ACC	Total No.
Condition 1	827	231	56	169	83.02%	93.65%	82.46%	1283
Condition 2	624	174	49	160	79.59%	92.71%	79.24%	1007
Condition 3	474	179	51	115	80.47%	90.28%	79.73%	819
All conditions	1925	584	156	444	81.26%	92.50%	80.70%	3109

TABLE 3. True positive rate (TPR), positive predictive value (PPV), and accuracy (ACC) of proposed algorithm on lip-reading recognition under individual differences.

	TP	TN	FP	FN	TPR	PPV	ACC	Total No.
Individuality 1	595	185	53	143	80.62%	91.82%	79.92%	976
Individuality 2	332	117	184	263	55.80%	64.34%	50.11%	896
Individuality 3	578	169	62	151	79.29%	90.31%	77.81%	960
Individuality 4	411	106	142	187	68.73%	74.32%	61.11%	846
All individualities	1916	577	441	744	72.03%	81.29%	67.78%	3678

and was classified into the silence and different vowel groups [17]. Table 1 shows the actual and predictive frames in all conditions. The actual frames correspond to vowels ‘a’, ‘i’, ‘u’, ‘e’ ‘o’ and silence were 248, 254, 299, 891, 593 and 824 respectively. The predictive frames correspond to vowels ‘a’, ‘i’, ‘u’, ‘e’ ‘o’ and silence were 237, 242, 327, 959, 556 and 788 respectively.

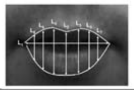
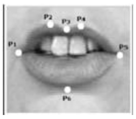
Table 2 showed the experimental results of vowel recognition under different conditions. It shows that the values of TPR, PPV and ACC of the proposed method for all conditions were 81.26 %, 92.50 % and 80.70 % respectively. From the experimental results, the recognition performances under the static and dynamic backgrounds were similarly good. Table 3 showed the experimental results of vowel recognition under individual difference. The values of TPR, PPV and ACC of the proposed method for all individualities were 72.03%, 81.29%, and 67.78% respectively. The influences of thick-lip makeup effect and 30-degree face angle on the recognition performance were unobvious. However, the factors of the moustache face and 60-degree face angle exactly reduced the recognition performance.

IV. DISCUSSIONS

In general, the current human lip recognition algorithms contain essentially two stages. First stage is to detect the possible location of the speaker’s face, and then to crop the part of the speaker’s mouth image [22]. The second stage is to parameterize the part of the mouth image by using geometric-based or image transform-based techniques. Petajan’s original system is one of geometric-based feature extraction that used simple thresholding of the mouth image to extract the lip area, and then measuring the information of mouth height, width and area [23]. Since then many approaches have also been developed to exploit the shape information of the human mouth to fit more complex lip models [24], [25]. In 2012, Ibrahim *et al.* proposed an automatic lip reading system to recognize the English digits (0-9) from the video sequences [26]. The tendency of several lip parameters extracted from continuous 17 frames in the video sequence was used to recognize English digits. Their experimental results showed that the recognition performance by using three lip parameters (lip height, lip width and lip ratio) was about 68%, and they also indicated that using more lip features (lip height, lip width, lip area, lip perimeter and lip ratio) could not improve the recognition performance. Moreover, the change of speech rate might affect the performance of this method on recognize English digits. In this study, several criteria for detecting the mouth region of interest were designed to reduce the influence of background image noise and other incomplete contours. Moreover, different from the above method, five lip parameters extracted from one frame in the video sequence were directly used to recognize English vowels when speaking. Therefore, the change of speech rate would not affect the performance of the proposed algorithm on lip-reading recognition. From the experimental results, the accuracy of the proposed algorithms on lip-reading recognition under varying background was over 80%. By using these lip parameters, the influence of the individual difference of the lip shapes, such as thick-lip makeup effect or 30-degree face angle, on the lip-reading recognition performance was unobvious. However, the moustache face and 60-degree face angle exactly reduced the recognition performance. Large face angle easily affected the detection of the mouth ROI. The moustache face might not only affect the detection of the mouth ROI, but also affected the size and position of the estimated lip contour, to reduce the recognition performance. Moreover, the individual difference of the speaking action might affect the performance; in particular, the change of the lip shape is extremely small when speaking. The occurrences of FP events were usually caused from that the lip feature of the specific vowel is similar to that of other vowel, for example, the lip feature of the vowel ‘e’ is similar to that of ‘i’, and that of ‘o’ is similar to that of ‘u’, or the unobvious change of lip shape when speaking.

Some approaches of moving human lip recognition have been proposed in previous studies, and the comparison between the proposed system and other methods was listed in Table 4. In 2004, Eveno *et al.* used the jumping snake approach for lip tracking [18]. Here, the user has to manually click a point as a seed near the vertical symmetry axis of mouth, and the pixels neighboring on the seed would be first

TABLE 4. Comparison between proposed method and other human lip-recognition algorithm.

	Wu *et al.* [7]	Cetingul *et al.* [10]	Liu *et al.* [11]	Eveno *et al.* [18]	Chin *et al.* [19]	Ibrahim *et al.* [27]	Proposed system
Used lip parameters	SDF and STLF	Six key points on outer lip contour, eight lip shape parameters 	12 normalized geometric features	Six key points on outer lip contour.	Six key points on outer lip contour 	Height, width and ratio of lip contour	Height, width, ratio, contour point and area of lip contour
Performance	TPR: 70.2 % @18.6dB, FPR: 7.1 % @18.6dB	ACC: 72.86 %, EER: 5.2%	TPR: 95.02%, FPR: 0.095%	Tracking error of key points: 4%	POL: 90% , SE: 12.6%	ACC: 68%	ACC: 80.70 %, TPR: 81.26%, PPV: 92.50%
Image source	Video frame (Oulu VS, PKU-AV)	Video frame	Video frame	TV news	Video frame (CUAVE, XM2TVS)	Video frame (CUAVE)	Video frame
Requirement of pretraining	Yes	Yes	No	No	No	No	No
Automatic mouth ROI detection	Yes	Yes	Yes	No	Yes	Yes	Yes
Applications	Audio-visual keyword spotting under noisy environment	Speaker Identification by lip contour and speech recognition	Speaker Identification by lip contour recognition	Lip contour recognition	Lip contour recognition	English digits recognition	English vowel recognition

evaluated to search one of current points on the lip contour and then detect the whole lip contour along this point by using the region-based active contour model (ACM). The total initialization time and total tracking time of this method on the computer with Pentium IV 2.4-GHz CPU were 0.8 seconds and 0.24 seconds respectively for 144 × 188 image (QCIF format). However, in this method, the location of the desired contour shape has to be selected manually. In 2006, Cetingul *et al.* utilized the explicit lip motion information associated with hidden Markov model (HMM) to do speaker identification and speech recognition [10]. In this study, the name scenario and digit scenario (a fixed 348–572 digit password in the digit dataset) were used for pre-training, i.e. the user has to utter ten repetitions of these sentences to pre-train this system. Here, the maximum horizontal distance, and the seven vertical distances from the Cupid's bow and from the equidistant upper lip points to the lower lip boundary were used as the features of the contour shape. The recognition rate (accuracy) of this method was about 72.86 %. Moreover, the requirement of this pre-training procedure also increased the inconvenience of use in daily life applications. In 2012, Liu *et al.* proposed a lip tracking approach [11]. In this method, the darkest sites in the grayscale photo were viewed as the mouth corners, and a relatively large value of divergence between neighboring pixels was recognized as contours. The correct accept rate (positive predictive value) and the false accept rate (false positive rate) of this method were greater than 95.02%, and smaller than 0.095% respectively. However, it easily failed when the darkest sites on image did not look like as the mouth corners or other environmental conditions affect, for example, black background. In 2012, a region-based active contour model with watershed segmentation was proposed for lips contour detection [19]. Here, the watershed algorithm, that was used to divide an image into several catchment basins of a heightmap and its outcome was the labeled pixels according to their respective catchment basin number, was used to segment a closed curve around the lip as the initial curve for the modified ACM function. The percentage of overlap (accuracy) of this method for CUAVE database and XM2TVS database were 90.0% and 86.2% respectively. In 2016, Wu *et al.* proposed a novel lip descriptor for audio-visual keyword spotting under noisy environment. In this study, the lip descriptor consists of the shape difference feature (SDF) and the spatiotemporal lip feature (STLF) to perform visual feature extraction. Moreover, an adaptive audio-visual integration strategy based on decision-level fusion proposed to make the best use of audio-visual speech and adapt to various noisy environments. Finally, 7 experimental results on the OuluVS and PKU-AV datasets demonstrated that the proposed lip descriptor shows superior performance. In 7th experiment, three kinds of noise intensity are estimated: weak noise with an average SNR of 18.6 dB, moderate noise with an average SNR of 11.2 dB and strong noise with an average SNR of 4.8 dB. In the condition of SNR of 18.6 dB, the TPR and false positive rate (FPR) are 70.2 % and 7.1 5 respectively. However, a pre-defined set of markers were required to divide an image. It might be difficult to satisfy different locations and sizes of the contour shapes. Moreover, the above methods could just detect the individual lip shape or check the opening or closing state of the mouth, but they could not precisely recognize each English vowel when speaking. For the proposed system, the total initialization time and the total tracking time on the computer with Pentium 2.94-GHz CPU were about 0.5 seconds and 0.27 seconds respectively. The accuracy of the proposed system on lip-reading recognition under different conditions was over 70 %. Different from other human lip recognition methods which could just detect the individual lip shape or check the opening or closing state of the mouth, the proposed method could recognize each English vowel when speaking. Moreover, the pre-training is not required for the proposed system, and this also improve the convenience of use.

V. CONCLUSIONS

A novel speech recognition algorithm was developed to extract the features of lip contour and real-time recognize each English vowel when speaking. Here, the criteria settings were developed to improve the stability of detecting the mouth ROI and lip contour. Five lip parameters extracted from one frame in the video sequence, including the width, height, contour points, area and the ratio (width/ height) of lips, were directly used to recognize English vowels when speaking. In this study, the total initialization time and the total tracking time were 0.5 seconds and 0.27 seconds respectively. Moreover, the mouth ROI could be detected automatically. Different from other human lip recognition algorithms, not only the lip contour but also English vowels could be recognized when speaking. The pre-training and the database for modeling individual lip contour were also not required, and this greatly improves the convenience of use in various kinds of application. From the experimental results, the accuracy of the proposed algorithm on lip-reading recognition was good (over 80%), and was insensitive to the varying background. The proposed method could automatically detect the mouth ROI, and reduce the influence of individual differences, such as the individual lip shape or makeup effect, and it also could performance a good performance without pre-training. This also improves its practicability and might contain the potential of applying in the lip-reading recognition under car driving in the future. However, reducing the brightness of the environmental background still obviously affected the performance of lip-reading recognition, and this is a problem to solve in the future.

REFERENCES

[1] A. P. da Silva, T. Feliciano, S. V. Freitas, S. Esteves, and C. A. E. Sousa, "Quality of life in patients submitted to total laryngectomy," *J. Voice*, vol. 29, no. 3, pp. 382–388, May 2015.

[2] N. Agrawal and D. Goldenberg, "Primary and salvage total laryngectomy," *Otolaryngol. Clin. North Amer.*, vol. 41, no. 4, pp. 771–780, Aug. 2008.

[3] A. Y. Chen and M. Halpern, "Factors predictive of survival in advanced laryngeal cancer," *Arch. Otolaryngol. Head Neck Surgery*, vol. 133, no. 12, pp. 1270–1276, Dec. 2007.

[4] C. E. Silver *et al.*, "Current trends in initial management of hypopharyngeal cancer: The declining use of open surgery," *Eur. Arch. Otorhinolaryngol.*, vol. 266, no. 9, pp. 1333–1352, Sep. 2009.

[5] A. Relic, P. Mazemda, C. Arens, M. Koller, and H. Glanz, "Investigating quality of life and coping resources after laryngectomy," *Eur. Arch. Otorhinolaryngol.*, vol. 258, no. 10, pp. 514–517, Dec. 2001.

[6] B. J. Borgstrom and A. Alwan, "A low-complexity parabolic lip contour model with speaker normalization for high-level feature extraction in noise-robust audiovisual speech recognition," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 6, pp. 1273–1280, Nov. 2008.

[7] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, "A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 326–338, Mar. 2016.

[8] S. L. Wang, A. W. C. Liew, W. H. Lau, and S. H. Leung, "An automatic lipreading system for spoken digits with limited training data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 12, pp. 1760–1765, Dec. 2008.

[9] J. Shin, H.-I. Kim, and R.-H. Park, "New interface for musical instruments using lip reading," *IET Image Process.*, vol. 9, no. 9, pp. 770–776, 2015.

[10] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 2879–2891, Oct. 2006.

[11] Y.-F. Liu, C.-Y. Lin, and J.-M. Guo, "Impact of the lips for biometrics," *IEEE Trans. Image Process.*, vol. 21, no. 6, pp. 3092–3101, Jun. 2012.

[12] X. Liu and Y.-M. Cheung, "Learning multi-boosted HMMs for lip-password based speaker verification," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 2, pp. 233–246, Feb. 2014.

[13] Z. Yi, L. Quan-Jie, L. Yan-hua, and Z. Li, "Intelligent wheelchair multimodal human-machine interfaces in lip contour extraction based on PMM," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Guilin, China, Dec. 2009, pp. 2108–2113.

[14] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Recognition of genuine smiles," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 279–294, Mar. 2015.

[15] C. Santiago, J. C. Nascimento, and J. S. Marques, "2D segmentation using a robust active shape model with the EM algorithm," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2592–2601, Aug. 2015.

[16] Y.-M. Cheung, M. Li, X. Cao, and X. You, "Lip segmentation under MAP-MRF framework with automatic selection of local observation scale and number of segments," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3397–3411, Aug. 2014.

[17] T. Song, K. Lee, and H. Ko, "Visual voice activity detection via chaos based lip motion measure robust under illumination changes," *IEEE Trans. Consum. Electron.*, vol. 60, no. 2, pp. 251–257, Jul. 2014.

[18] N. Eveno, A. Caplier, and P.-Y. Coulon, "Accurate and quasi-automatic lip tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 706–715, May 2004.

[19] S. W. Chin, K. P. Seng, and L.-M. Ang, "Lips contour detection and tracking using watershed region-based active contour model and modified H_∞," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 6, pp. 869–874, Jun. 2012.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Kauai, HI, USA, Dec. 2001, pp. I-511–I-518.

[21] O. Etesami and A. Gohari, "Maximal rank correlation," *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 117–120, Jan. 2016.

[22] D. Stewart, A. Pass, and J. Zhang, "Gender classification via lips: Static and dynamic features," *IET Biometrics*, vol. 2, no. 1, pp. 28–34, Mar. 2013.

[23] E. D. Petajan, "Automatic lipreading to enhance speech recognition (speech reading)," Ph.D. dissertation, Dept. Elect. Eng., Univ. Illinois Urbana-Champaign, Champaign, IL, USA, 1984.

[24] R. Kaucic, B. Dalton, and A. Blake, "Real-time lip tracking for audio-visual speech recognition applications," in *Proc. Eur. Conf. Comput. Vis.*, Cambridge, U.K., 1996, pp. 376–387.

[25] M. Gordan, C. Kotropoulos, and I. Pitas, "Pseudoautomatic lip contour detection based on edge direction patterns," in *Proc. IEEE Conf. Image Signal Process. Anal. (ISPA)*, Pula, Croatia, Jun. 2001, pp. 138–143.

[26] M. Z. Ibrahim and D. J. Mulvaney, "Geometry based lip reading system using multi dimension dynamic time warping," in *Proc. IEEE Conf. Vis. Commun. Image Process. (VCIP)*, San Diego, CA, USA, Nov. 2012, pp. 1–6.

BOR-SHING LIN (M'11) received the B.S. degree in electrical engineering from National Cheng Kung University, Taiwan, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taiwan, in 1999 and 2006, respectively. Since 2009, he has been a member of the Faculty with the Department of Computer Science and Information Engineering, National Taipei University, Taiwan, where he is currently an Associate Professor. His research interests are in the areas of embedded system, wearable device, IoT, biomedical signal processing, biomedical image processing, portable physiological monitoring system, and rehabilitation engineering.

YU-HSIEN YAO received the M.S. degree in imaging and biomedical photonics from National Chiao Tung University, Hsinchu City, Taiwan, in 2016. His current research interests are in the areas of digital image processing and biomedical signal processing.

CHING-FENG LIU is currently with the Department of Medical Research, Chi Mei Medical Center, Tainan, Taiwan, and also with the Graduate Institute of Medical Sciences, Chang Jung Christian University, Tainan City, Taiwan.

CHING-FENG LIEN is currently with the Department of Otolaryngology–Head and Neck Surgery, E-Da Hospital, Kaohsiung, Taiwan, and also with I-Shou University, Kaohsiung, Taiwan.

BOR-SHYH LIN (M'02–S'15) received the B.S. degree from National Chiao Tung University, Taiwan, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taiwan, in 1999 and 2006, respectively. He is currently a Professor with the Institute of Imaging and Biomedical Photonics, National Chiao Tung University. His research interests are in the areas of biomedical circuits and systems, biomedical signal processing, and biosensors.