International Journal of Innovative Computing, Information and Control
Volume **8**, Number **4**, April **2012**

ICIC International ©2012 ISSN 1349-4198
pp. 2837–2853

A ROBUST AND REAL-TIME VISUAL SPEECH RECOGNITION FOR SMARTPHONE APPLICATION

Min Gyu Song[1], Md Tariquzzaman[1], Jin Young Kim[1]
Seong Taek Hwang[2] and Seung Ho Choi[3]

[1]School of Electronics and Computer Engineering
Chonnam National University
300 Yongbong Dong, buk-Gu, Gwangju 500757, South Korea
tareq@moiza.chonnam.ac.kr; smg686@icdsp.jnu.ac.kr; beyondi@jnu.ac.kr

[2]Communication Research Center
Multimedia Lab, IT Center
Samsung Electronics
Suwon, Gyeonggi-do 442600, South Korea
shwang@samsung.com

[3]Department of Computer Science
Information Center
Dongshin University
Kunjae-Ro, Naju, Chonnam 520714, South Korea
shchoi@dsu.ac.kr

Received November 2010; revised March 2011

Abstract. *Visual speech recognition (VSR) is one prospective complementary approach for speech recognition under very noisy environments, especially in mobile phone circumstances. In implementing visual speech recognition on a smartphone, the two main issues of real-time responsiveness and robustness conflict with each other. In this paper we proposed and implemented a robust visual speech recognition system that performs in real-time. First, we devised a robust and fast lip detection method based on eye-detection, which is not vulnerable to changes in illumination. The pair of eyes was determined based on image binarization and a coupled-eye validation method. Then the lip region was estimated by geometric lip candidate detection and k-means clustering. Second, to cope with the problem of lighting-dependent visual speech recognition performance, we combined the previous methods of lip-folding and RASTA filtering and introduced a modified histogram equalization, in which a mapping function was calculated for the first frame image and fixed through the following images. Third, the visual speech recognition system with 32 control words was implemented on a smartphone with code optimization. It was shown to work in real-time with promising results.*

Keywords: Visual speech recognition, Lip localization, K-means clustering, Histogram matching, Lip folding, RASTA filter

1. **Introduction.** With the advent of smartphones in our mobile society, the importance of natural human machine interaction (HMI) is continuously increasing. However, state-of-the-art smartphones still adopt conventional keypads or touchpads for HMI. The touchpad is replacing the keypad or mouse due to the small size of mobile phones. There is no doubt that automatic speech recognizers (ASRs) are installed on smartphones. Nevertheless, ASRs have not been fully used, as typical ASR performance is severely degraded when used in noisy environments. Therefore, ASR has limited commercial applications in games, door locks, call-centers, or remote controls. To cope with the problem of ASR performance degradation in noisy surroundings, three broadly different approaches have

been proposed in the literature: noise cancellation, noise robust feature extraction, and model adaption to environments [1-10]. However, these methods have achieved limited success when applied to real environments.

On the other hand, in 1976, H. McGurk and J. MacDonald showed that there was an interaction between hearing and vision in speech perception [11]. After the demonstration of the McGurk effect, studies on automatic lip reading or visual speech recognition have been attempted as supplementary methods to audio-only speech recognition by many researchers. In audio-visual speech recognition (AVSR), the main topics of study are in the development of a reliable VSR system and a proper fusion strategy. Since 1990, extensive studies have been performed on such topics. These studies considered only acoustic noise, assuming that the visual signal did not experience any distortion [12-16]. However, in real service environments, there are different sources for visual signal distortion: illumination changes, miss-detected lips, slanted lips, rotated lips, and so on. Thus the robustness of a VSR system has been disregarded compared with ASR systems.

In this paper, we propose an illumination-robust and fast VSR scheme which works in real time on a mobile phone. Basically, our VSR system is composed of lip detection or localization, lip feature extraction, and classification of the featured stream into specific words. Undoubtedly, precise lip localization (LL) and robust lip feature extraction (LFE) play an important role in the VSR system.

To achieve lighting-robust lip detection, we proposed an eye-detection based lip localization without complete face detection. This is because the face color space is easily changeable relating to illumination conditions, whereas the color space of the eye pupils is relatively stable regardless of the lighting conditions. After eye localization, we framed a box that included the lips by using the geometric characteristics of the face. Then, the precise lip region was decided by k-means clustering. The clustering approach automatically sets a threshold to detect the candidate regions of the dark inner lip.

In our VSR system, we tested and verified the previous methods – lip folding and RASTA filtering [17] – with DCT and PCA, developed for robust lip-reading. To make the system more robust, we proposed a modified histogram matching function and incorporated it into the conventional approaches. Our suggestion was that the mapping function of the histogram equalization was estimated using only the first frame, the image of which showed closed lips. After computing the mapping function, this function was kept through the last frame images.

The whole system was evaluated on a PC and then implemented on a smartphone. After code optimization, we confirmed the real-time functionality of our VSR system.

The organization of this paper is as follows. The overall eye tracking method and lip localization with the proposed approach are described in Section 2. In Section 3, the proposed approach for robust feature extraction is detailed. Experimental results and the implementation of the proposed VSR on a smartphone platform are detailed in Section 4. We conclude this paper in Section 5.

2. Lip Localization Based on Eye Detection.

Precise lip localization in video is an active research field in computer vision. Certainly, if a lip-reading system fails to find the location of the lips, it causes a significant distortion of the visual features and thus leads a severe degradation in VSR performance. Hence, lip localization is a decisive preprocessing for a reliable VSR.

The conventional approach for detecting the lips in an image is to use color information [18-21]. In these methods it is assumed that the image includes only the lip region or the face region. The face region is determined before lip detection. These existing approaches work fine under controlled illumination. However, mobile environments are full of dynamic

lighting conditions. In addition, every speaker has a different lip color. Moreover, in the color-based approach, threshold selection to distinguish lip color and skin color often results in errors in lip detection.

In this paper we proposed a lip localization scheme using eye detection. The pupils are the darkest object in a speaker's face and are not highly variable under illumination compared with other parts of the face. Thus, the eyes could be detected regardless of the lighting conditions. After finding a set of eyes we can easily guess a rough region that includes the lips using geometric information between facial objects, i.e., the eyes and lips. Figure 1 shows the schematic diagram of the proposed lip detection approach. Briefly, a binary image is taken from an input image in order to navigate to the eye candidates using an adaptive threshold.

After the adaptive threshold, a GMM validation is performed on the candidate segments of the pair of eyes including the eyebrows. Then, based on the detected location of the eyes, a rough lip region is determined. From the rough lip region, the smallest right rectangle that includes the lips is obtained through k-means clustering [22].

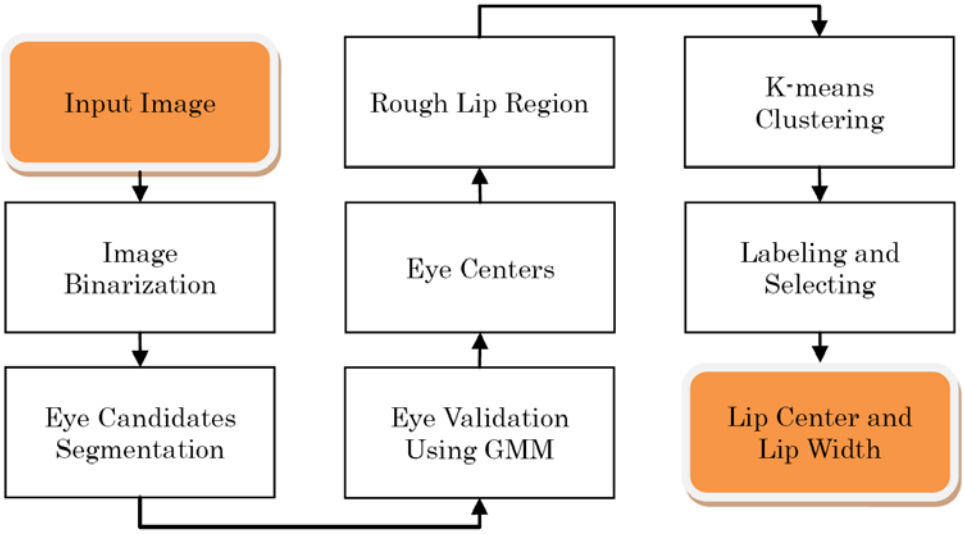


FIGURE 1. Lip detection flowchart

2.1. **Eye segmentation based on adaptive thresholding.** Typically the color of the pupil is darker than the surrounding face colors. This color information in a facial image is a clue for an intelligent machine to locate the eyes. For a mobile application of a prospective VSR system, a frontal facial image is taken near the center of the screen. The property of the color variation in the facial image and its location on the screen is the place where the application finds the rough eye region, where it applies a mask. The masked image is then converted into a binary image based on the adaptive thresholding technique. The binarization is performed on the Y (luminance) space after transforming the RGB image to YCbCr. We applied two kinds of adaptations: an average luminance-based adaptation (ALBA) and a candidate number-based adaptation (CNBA). Basically, ALBA is used to set the initial threshold by:

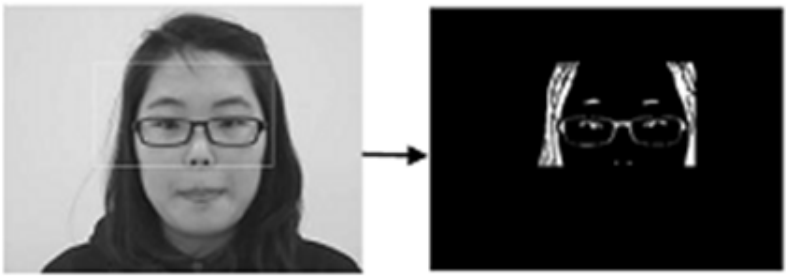
$$ThrE = f(Ymed) = 0.3612Ymed + 25.06 \tag{1}$$

where $ThrE$ is the threshold for eye segmentation and $Ymed$ is the median luminance value of the central input image region, which is occupied only by the face in pixels. The linear model parameters are determined by the curve fitting of the experimental data. Each pixel of the luminance image is converted into one if the particular pixel intensity is

less than the adaptive threshold value. Otherwise, the pixel is set to zero. Figure 2 shows an example of the results of the threshold process.

Adaptive eye segmentation is good enough to crop the eye region. However, there still might be significant non-eye regions in the crop. These regions should be removed for precise eye region extraction. The straightforward process for validation of each eye region is based on the physical characteristics of the eye. In the simple validation process, the regions which have a height greater than 3 times the width are rejected. Figure 3 shows the binary image after object validation.

On the other hand, the ALBA approach does not guarantee that at least one candidate pair of eyes still exists after the geometric validation. In that case we have to increase the threshold until the proper candidates are found which we term the CNBA. Figure 4 shows the schematic diagram of the CNBA.



(a) (b)

FIGURE 2. (a) The input Y image, (b) the eye candidates regions

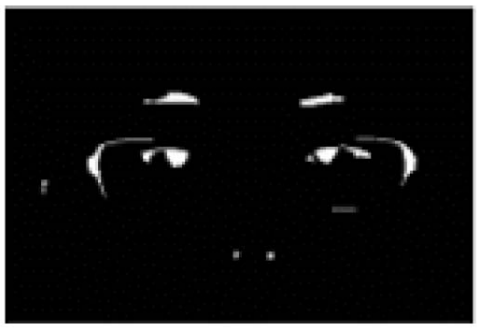


FIGURE 3. The remaining eye candidates

2.2. **Eye validation based on GMM.** To validate each candidate object, we applied the GMM method. Generally, each object is evaluated separately. An alternative validation unit for a single object is a coupled object of the left and right eye candidates. We can expect that the coupled object approach enhances the validation performance, because the left and right eyes are tightly correlated. Also, eyebrows are good evidence for eyes below. Therefore, we extended the coupled object so that eyebrow regions are included. Figure 5 shows each validation unit.

For the given validation images, an image-size normalization is performed first. Then, we use a Haar wavelet transform approach for feature extraction from the eye image. PCA is applied to these features in order to reduce the feature dimensions to 8. For GMM-training with 3 mixtures we used 1,000 training images for all cases. For the testing objects, we calculated the probabilities of each object using the GMM probability density

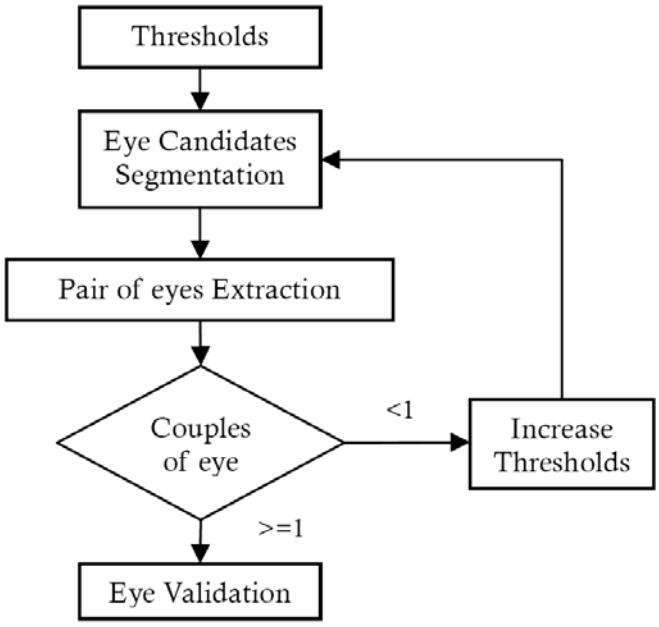
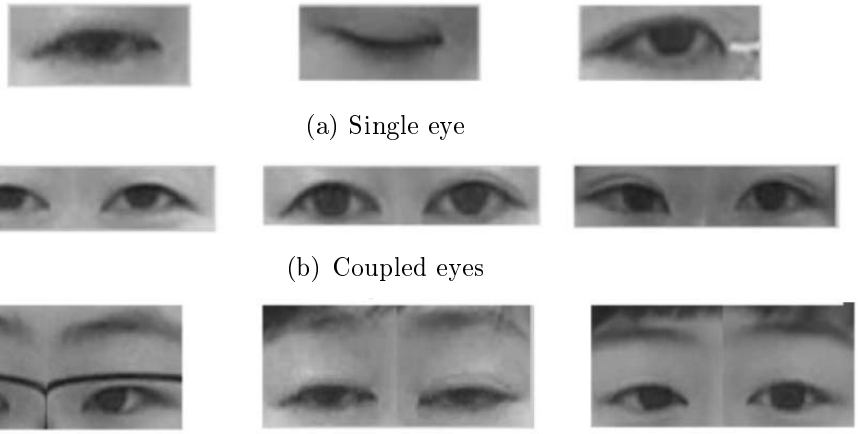


FIGURE 4. The adaptive segmentation algorithm (CNBA)



(a) Single eye

(b) Coupled eyes

(c) Coupled eyes with eyebrows

FIGURE 5. The eye validation units

TABLE 1. Comparison of the different validation units

	Standard database	Indoor database
Single eye	60%	–
Coupled eyes	90.3%	87.2%
Coupled eyes with eyebrows	98.1%	96.6%

function. The objects with the maximum probability are selected as the eyes. The overall eye validation procedure is shown in Figure 6.

We used 1000 test images to evaluate the proposed eye validation approach. We created a database specifically for training and testing. The specifics of the database are described in Section 4. Table 1 shows the comparison results of the different validation units for the standard database and the indoor database.

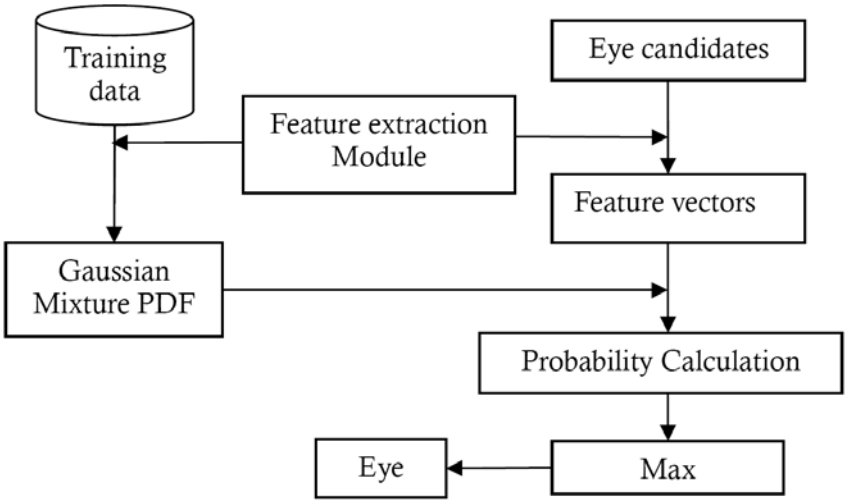


FIGURE 6. The eye detection flowchart based on GMM

From Table 1 we observed that using the validation unit of the coupled eyes with eyebrows showed the best results regardless of the lighting conditions.

2.3. **Geometrical relation-based rough lip region segmentation.** Faces have some natural symmetry. For example, the left and right eyes are symmetrical on the vertical axis that usually goes through the nose and mouth. The distance between the two eyes is approximately equal to that of a lip's width. Also, the outer distance approximately equals the distance from the midpoint between the eyes to the center of the mouth.

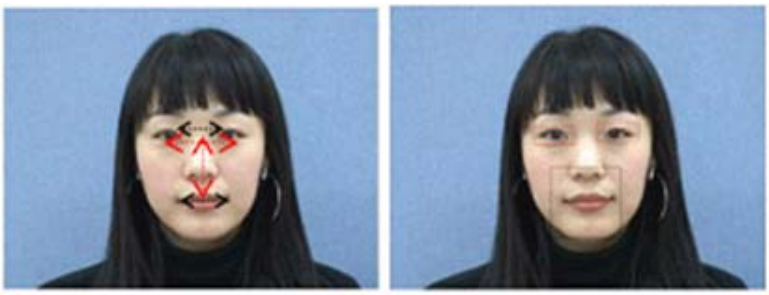


FIGURE 7. The rough cropped-lip region using the geometrical distance relationship of eyes and mouth

Thus we use these characteristics to determine the potential region of the mouth. Figure 7 shows a typical rough lip region found using the geometrical approximation method. This region may also contain a nose, a chin, shadows, and the lower parts of the face. Therefore, we need a precise and valid lip region for reflecting proper lip motion in speech recognition. Hence, we have applied the k-means clustering algorithm which is detailed in the following section.

2.4. **Precise lip region detection.** In our VSR system, we extracted lip features using an image transform-based method. Thus, the purpose of exact lip region detection is to set a box of the right rectangle around the lip center. The width of the box is the same as the lip width. The main aims are to obtain the lip center and to find the lip width. Our idea is that the inner lip region looks comparatively dark without being affected by the lighting conditions. To avoid setting a hard threshold we apply a k-means clustering

algorithm to detect the dark regions contained in the rough lip region. The process is as follows:

1) Perform k-means clustering for all of the pixels in the rough lip region. The feature vector is the RGB color values of each pixel.

2) Select the darkest clusters having the lowest intensity value among the centroid vectors.

3) Label the clustered image and select the darkest objects with the largest size.

To reduce the computational time for the clustering, the lip image candidate is down-sampled by a ratio of 2:1. The k-means clustering algorithm is applied to the compressed image. The k value was selected experimentally as 4 while dealing with the three-dimensional RGB image. Figure 8 shows the block diagram for obtaining the lip candidate region.

Among the inner lip candidates, we selected the best one according to size (number of pixels in the object). Based on the largest selected object we calculated the left and right boundary points and the lip center. Usually, the central point is obtained by measuring the gravity of the mass of the object. However, this can lead an error if the largest object in the clustered image does not have right and left symmetry due to illumination conditions. Therefore, we obtained the left and right end points using the width of the largest object of the clustered lips image. Figure 9 shows the selected object using the width of largest object in the clustered image, and the marker shows both sides of the lips.

Figures 10 and 11 show some of the samples of the lip localization approach in the respective facial image in standard and indoor environments, respectively. The proposed approach has showed its effectiveness in both environments to be able to collect the lip region, which is essential to reflect visual motion through feature characteristics needed in speech recognition.

3. **Robust Feature Extraction.** In this paper, different feature extraction approaches are tested to get the robust visual features needed to improve the recognition performance of the VSR system. There are some state-of-the-art features that are robust

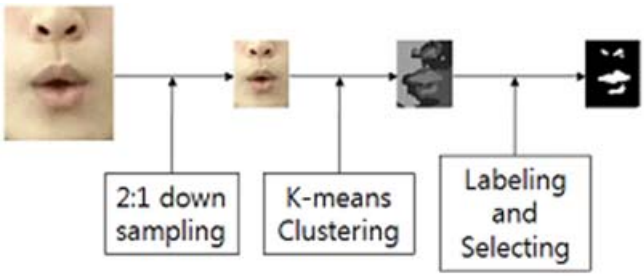


FIGURE 8. The extraction of the lip candidate objects

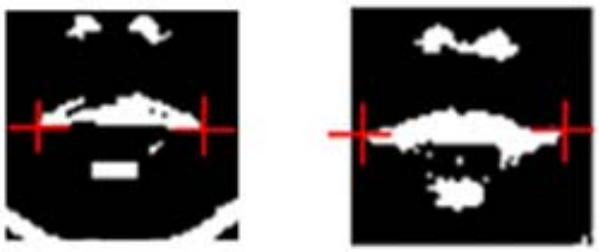


FIGURE 9. Lip corner detection

FIGURE 10. Lip detection in a standard environment

FIGURE 11. Lip detection in an indoor environment

in any illumination condition [17,23,24]. Generally visual speech feature extraction approaches are generally classified into two categories: the transform-based approach and geometrical feature-based approach. Previous studies have shown that transform-based approach performance is better than geometrical features-based approach [25]. Therefore, the transform-based approach was implemented in this paper. The basic steps for the method are as follows:

1) Perform image normalization into 32×32 pixels.
2) Apply DCT to transform image into the frequency domain.
3) Apply PCA to reduce the feature dimensions.

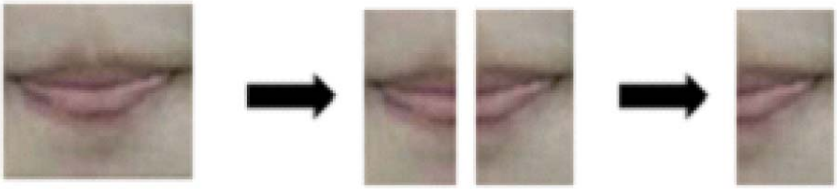


FIGURE 12. Lip folding based on geometrical symmetry

Additionally, we adopted lip folding and inter-frame filtering, which is proposed in Kim et al. [17]. These approaches were already evaluated for natural lighting conditions and were proven to deliver good preprocessing to enhance lip-reading performance. Lip folding is based on symmetry between the left and right halves of the lips. The effect of lip folding is to reduce the impact of the illumination mismatch of the left and right regions. Figure 12 shows lip folding based on geometric symmetry.

Inter-frame filtering originated from RASTA filtering in speech recognition. Band-pass filtering is performed for image sequences. Actually, the high-pass and low-pass filtering processes are performed sequentially. The filters are as follows:
High-pass filter:

$$Y_t[n,m] = 0.9859 \times (X_t[n,m] - X_{t-1}[n,m]) + 0.9716 \times Y_{t-1}[n,m] \tag{2}$$

Low-pass filter:

$$Y_t[n,m] = 0.8638 \times (X_t[n,m] - X_{t-1}[n,m]) - 0.7257 \times Y_{t-1}[n,m] \tag{3}$$

To achieve a higher performance of the VSR, we adopted a histogram matching approach. We tried to enhance the recognition rate using histogram equalization. However, the experimental results did not show a good performance. As an alternative method, we made use of histogram matching (HM) [26]. For histogram matching we needed the reference or target histogram of some lip images. This target histogram was obtained by averaging the histograms of all the training lip images. For a given lip image, the histogram matching function was estimated by comparing the target histogram and the histogram of the input image. Figure 13(a) shows the standard or target histogram and a sample image histogram and Figure 13(b) shows an example of the matching function.

In this paper, we suggested two HM applications for illumination compensation. The HM was performed before the lip-folding process. The second method was very commonly used in illumination compensation.

■ ***Method 1 (HM0)*** – For all lip images in the lip sequences the matching function is calculated and applied for each lip image separately.

■ ***Method 2 (HM1)*** – The matching function is calculated only for the initial or first lip image and applied to the last lip sequence images.

DCT is used as a last step for feature extraction and PCA is applied to obtain the reduced lip features in the VSR scheme. In this study, we use two-dimensional DCT on a 32×16 image, resulting in the DCT features. For our experiment we took only the low frequency components of the 16×8 section discarding the high-frequency domain. Finally, the PCA was applied to get the lip feature. The schematic procedure of the lip feature extraction is shown in Figure 15. In our experiments, 4 principal vectors were used for feature dimension reduction. Also, we added 4 delta parameters, so the feature dimension is 8. Delta parameters were calculated simply as follows:

$$\mathbf{d}[n] = (\mathbf{x}[n+1] - \mathbf{x}[n-1])/2 \tag{4}$$

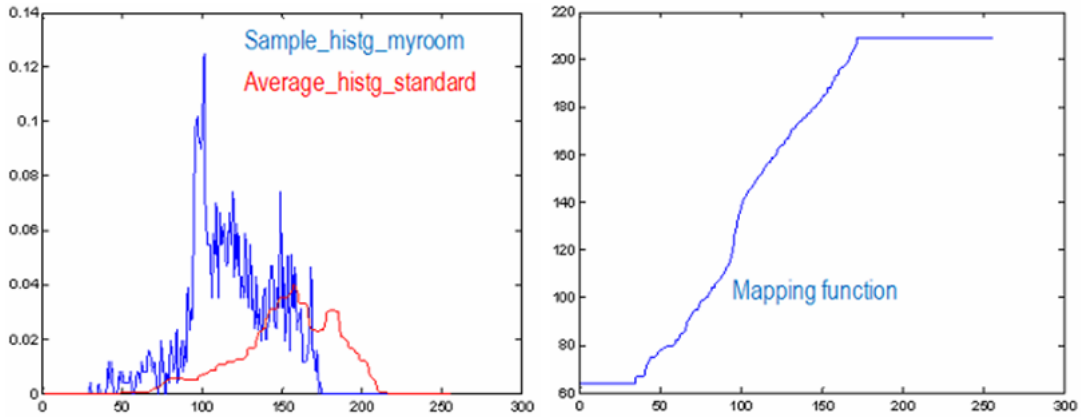


FIGURE 13. The histogram matching process: (a) target and input histograms, (b) mapping function between target and input histograms

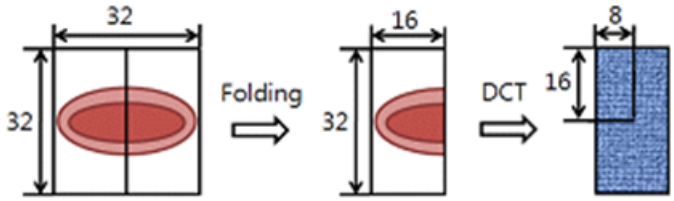


FIGURE 14. An example of the DCT window

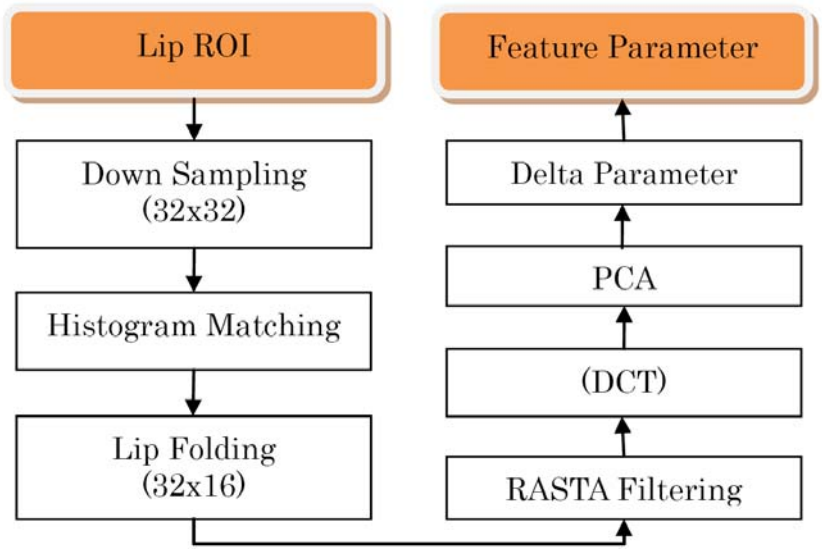


FIGURE 15. The schematic diagram for the lip features extraction

where $\mathbf{x}[n]$ is the 4-dimensional feature vector for the nth frame and $\mathbf{d}[n]$ is the corresponding delta parameters.

4. Experiment and Analysis.

4.1. **The experimental database.** We constructed an audio-visual database of 110 speakers and 32 isolated smartphone control words using a SONY digital camcorder. All subjects spoke the 32 words two times under two different lighting conditions. The

speakers of the collected database were 20-40 years old. The main specs of our AV database are shown in Table 2. The database was collected in different environments over a period of three months. Depending on the recording environments, the databases were classified into two separate groups: the standard database and the indoor database. The standard database meant that recording was done in a lighting-controlled room, that is, during the collection of the standard database the lighting conditions were uniform and fixed. The indoor database had a variety of backgrounds and lighting condition changes in comparison to the standard database. Figures 16 and 17 show example images of the standard database and indoor database, respectively.

The 32 words were grouped into four as shown in Table 3. In Table 3, we show the Korean pronunciation in parentheses with the corresponding English words. Groups 1 and 2 included the words which were chosen to command the music player device. Group 3's words were chosen to inform the user context, while Group 4's words were selected as decisive commands.

TABLE 2. Constructed database

Database	Definition	Number of Words	Number of Subjects	Capture Specs
Standard database	Recorded in illumination-controlled condition	32 (smartphone control words)	118 (Number of males: 53 Number of females: 65)	Fs = 30Hz Image size = 640 × 480
Indoor database	Recorded in various office conditions			

TABLE 3. Korean word group information which includes words written in the English alphabet and English-language translations

Word Group Unit	Words in Group
Word Group 1	재생-jaesaeng(PLAY), 정지-jeongji(STOP), 일시정지-ilsijeongji(PAUSE), 다음곡-daeumgok(NEXT SONG), 이전곡-ijeongok(PREVIOUS SONG), 음악정보-eumakjeongbo(FILEINFOMATION), 소리크게-sorikeuge(VOLUMUP), 소리작게-sorijakge(VOLUMEDOWN)
Word Group 2	Play, Stop, Pause, Next, Previous, Fileinfo, Volumeup, Volumedown
Word Group 3	회의중-hoeuijung(UNDER MEETING), 운전중-unjeonjung(UNDER DRIVING), 수업중-sueopjung(IN THE CLASS), 도서관-doseogwan(IN THE LIBRARY), 독서실-dokseosil(IN THE READINGROOM), 영화관-yonghwagwan(IN THE THEATER), 도와줘-dowajwo(HELP ME),SOS
Word Group 4	네-ne(YES), 아니오-anio(NO), YES, NO, 확인-hwakin(OK), 취소-chwiso(CANCEL), OK, CANCLE

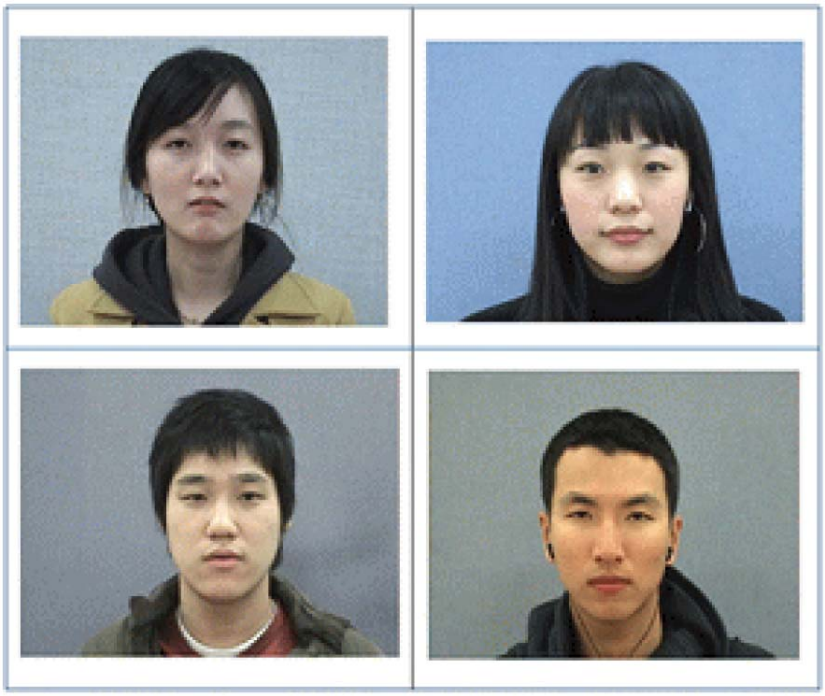


FIGURE 16. An example of the standard database

FIGURE 17. An example of the indoor database

4.2. **Experimental results.** The classical HMM-based visual speech recognition experiments were carried out with the AV database based on the different approaches, either in a single or in a combination manner, for the features extraction described above. The recognition experiment was performed using the following group of datasets:

■ Training data:

DB Group I: Speakers 1 to 80 from the standard database.

■ Test data:
DB Group II: Speakers 81 to 110 from the standard database: standard and inter-speaker (***SInter***).
DB Group III: Speakers 81 to 110 from the indoor database: indoor and inter-speaker (***IInter***).
DB Group IV: Speakers 1 to 80 from the indoor database: indoor and intra-speaker (***IIntr***).

DB Groups II and III are composed of the same speakers. While the speakers of DB Group IV are the same as the training speakers, the database of DB Group III was collected in different illumination conditions, i.e., an indoor environment. Table 4 shows the VSR experimental conditions.

TABLE 4. VSR experimental conditions

Frame Rate	Lip Features	Feature Dimensions	Recognition Method
30Hz	Image-transform based feature PCA + (FOL + FIL + DCT + HM0(1))	4 PCAs + 4 delta parameters.	Hidden Markov model Number of state = 1.5∗ (num of phonemes in a word) Number of mixture = 3
ORL: Original (Basic) 32 × 32 image, FOL: 32 × 16 Folded image, FIL: Inter-frame (RASTA) Filtered images, DCT: Discrete Cosine Transform, PCA: Principal Component Analysis, HM0: Typical Histogram Matching, HM1: Proposed Histogram Matching			

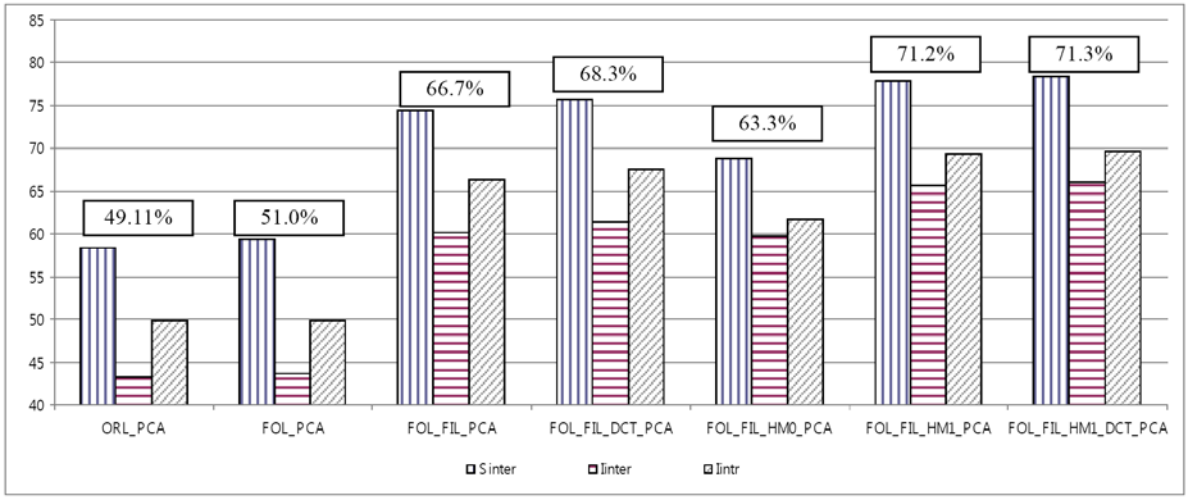


FIGURE 18. The comparison of the recognition rate (%) using the different feature extraction approaches (* the overall average VSR rates are shown in the rectangles)

The experimental results are shown in Figure 18. It is observed in Figure 18 that with the original common feature using only PCA and delta-parameters the performance average of 49.1% was obtained considering all the word groups and lighting conditions. We could confirm that adopting lip folding and inter-frame filtering enhanced the recognition rate to 68.3%. The best performance, an average of 71.3% for all the testing conditions, is achieved in the case of FOL + FIL + HM1 + DCT + PCA. Thus, the proposed histogram-matching method (HM1) achieves 3% error reduction. But, HM0 decreases the recognition rate from 68.3% to 63.3%.

Table 5 shows the FOL + FIL + HM1 + DCT + PCA approach in detail for each word group and each database group. From the experimental results it was observed that incorporating the DCT did not improve the recognition rate significantly in comparison with the recognition performance of FOL + FIL + HM1 + PCA. When comparing all the experimental results, it was seen that the lip folding approach, inter-frame filtering, and histogram matching (HM1) techniques were promising in VSR along with the existing state-of-the-art feature extraction approach.

TABLE 5. Each group performance for FOL + FIL + HM1 + DCT + PCA

Word. Gr. / DB Gr.	Word Group 1 Avg. (%)	Word Group 2 Avg. (%)	Word Group 3 Avg. (%)	Word Group 4 Avg. (%)	Overall Avg. (%)
Sinter	82.1	85.3	77.1	69.2	78.4
IInter	69.2	71.8	64.9	58.8	66.1
IIntr	72.5	75.2	68.1	61.6	69.4

4.3. **Implementation in the smartphone environment.** The entire VSR process was implemented and validated in a smartphone environment. The smartphone specifications were: 806 MHz CPU, 128 MB RAM, and 256 MB ROM. The image captured by the front camera of the smartphone was 320×240 pixels in the form of an RGB color model with a sampling rate of 15 frames/second. Figure 19 shows some of the scenarios used while implementing the VSR system in the smartphone. Figure 19(a) shows the initial display of our implemented system. In the implemented system, the black box was used as a suggestive eye location guideline to capture the highly tuned visual speech information in order to have a more precise lip region. Figure 19(b) shows the word group selection among the different action-based word groups. When we selected G2, the smartphone showed all the action words in Word Group 2. Then the VSR process was performed as follows:

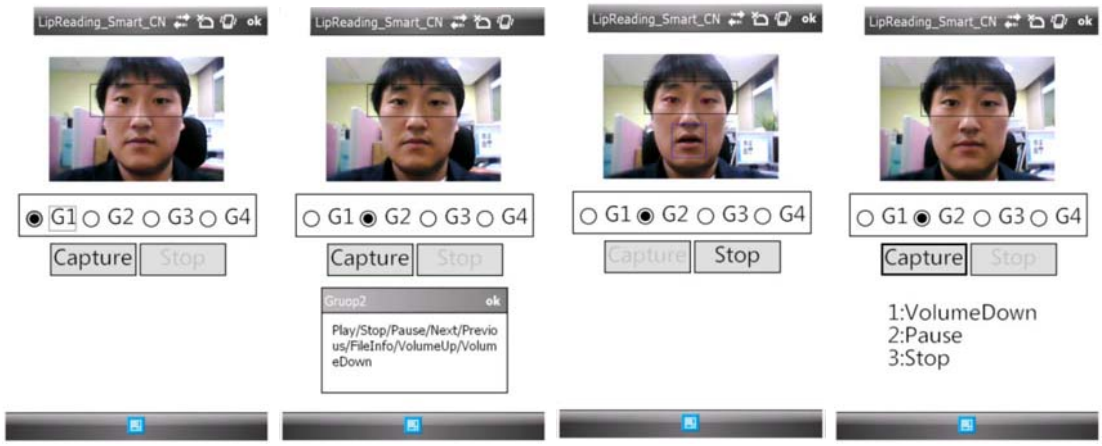
1) As soon as a user pressed the capture button, the VSR system tried to find the eye positions for the first captured frame. After eye localization the process displayed a sign, which initiated a user to speak.

2) While a user spoke, the system tracked the lip regions and analyzed the lip image sequences. In this stage lip tracking was performed based on the previous lip locations, so eye detection did not work in this stage. Figure 19(c) shows the capture process when a speaker uttered the respective action-based word.

3) When the user pressed the stop button, the system ended the image capture process. Then the system performed the recognition process of HMM and displayed the recognition results. Figure 19(d) shows the final stage of the implemented system with *a priori* order. In Figure 19(d) the user uttered "Volume Down", and the system displayed "VolumeDown" as highest priority word on its screen.

According to the VSR smartphone recognition experiments, the most time-consuming module of the smartphone's VSR system was eye detection. The eye-detection based lip localization was performed during the first frame as described in the scenario. With the online VSR experiments, we found that the eye-detection process took 0.5-1 sec. However, users did not feel any response delay in the lip tracking and recognition process.

The overall average recognition performance of the smartphone-based VSR indoors for each group of words is summarized in Table 6. The selections of the speakers were done the same way as in test Word Group 3 for real-time VSR operation. It is seen from Table 6 that smartphone performance is similar in comparison with PC-based experimental results. However, there were some deviations between the two results which are presented in Tables 5 and 6. These deviations were basically due to smartphone user error while



(a) (b) (c) (d)

FIGURE 19. The implementation of the visual speech recognition system in a smartphone environment

TABLE 6. Each group's performance on the smartphone

Group Unit	Word Group 1 Avg. (%)	Word Group 2 Avg. (%)	Word Group 3 Avg. (%)	Word Group 4 Avg. (%)	Overall Avg. (%)
Recog. Rate	68.3	73.7	65.3	58.3	66.4

handling the handset. With current state-of-the-art smartphones it is easy to focus the eye onto the suggestive guidelines particularly in the visual word recognition application domain. Nevertheless, during the uttered visual word acquisition through the smartphone for testing in a mobile environment, a very few speakers' hands shook or were unable to fix their eye on the suggestive guideline through the smartphone. Specifically, these two reasons created some errors in smartphone-based word recognition which could be considered constraints of user input detection in handling VSR through a smartphone. But nevertheless the proposed approaches for real-time VSR performance on smartphones are promising.

5. **Conclusion.** In this paper, we developed a VSR scheme for implementing a real-time VSR on a smartphone. For robust lip detection, we proposed a lip region detection method beginning with eye detection. Lip folding, inter-frame filtering and histogram matching approaches were applied to get robust features under dynamic illumination changes. The primary system was constructed and tested on a desktop PC. The evaluated VSR system was then implemented on a smartphone as a perspective. Through VSR testing with the smartphone we confirmed that the developed VSR system worked satisfactorily compared with the desktop version.

In this work we only dealt with visual signal-based speech recognition. In future work, we will focus on integrating the visual modality with the audio modality to ensure the recognizing system is robust in both noisy environments and variable lighting conditions. The current developed system has two limitations: adopting the guiding box on the smartphone implementation and the time-consumption of eye detection. So we will develop an unconscious guiding method and cope with time problem of eye detection.

Acknowledgment. This research was supported by the Korean Ministry of Knowledge Economy under the Information Technology Research Center support program supervised by the National IT Industry Promotion Agency (NIPA-2010-C1090-1011-0008) and Samsung Electronics.

REFERENCES

[1] S. Furui, Cepstral analysis technique for automatic speaker verification, *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol.29, no.2, pp.254-272, 1981.

[2] B. S. Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *J. Acoust. Soc. Am.*, vol.55, no.6, pp.1304-1312, 1994.

[3] T. Guan and Q. Gong, A study on the effects of spectral information encoding in mandarin speech recognition in white noise, *ICIC Express Letters*, vol.3, no.3(A), pp.415-420, 2009.

[4] H. Hermansky and N. Morgan, RASTA processing of speech, *IEEE Trans. on Speech and Audio Processing*, vol.2, no.4, pp.578-589, 1994.

[5] X. Wang, J. Lin, Y. Sun, H. Gan and L. Yao, Applying feature extraction of speech recognition on VoIP auditing, *International Journal of Innovative Computing, Information and Control*, vol.5, no.7, pp.1851-1856, 2009.

[6] R. Chengalvarayan and L. Deng, A maximum a posteriori approach to speaker adaptation using the trended hidden Markov model, *IEEE Trans. on Speech and Audio Processing*, vol.9, no.6, pp.549-557, 2001.

[7] C. H. Sit, M. W. Mak and S. Y. Kung, Maximum likelihood and maximum a posteriori adaptation for distributed speaker recognition systems, *Proc. of the 1st International Conference on Biometric Authentication*, pp.640-647, 2004.

[8] C. Kim and K.-D. Seo, Robust DTW-based recognition algorithm for hand-held consumer devices, *IEEE Trans. on Consumer Electronics*, vol.51, no.2, 2005.

[9] R. Flynn and E. Jones, Robust distributed speech recognition using speech enhancement, *IEEE Trans. on Consumer Electronics*, vol.54, no.3, pp.1267-1273, 2008.

[10] Y. Lu et al., Robust speech recognition using improved vector taylor series algorithm for embedded systems, *IEEE Trans. on Consumer Electronics*, vol.56, no.2, 2010.

[11] H. McGurk and J. MacDonald, Hearing lips and seeing voices, *Nature*, vol.264, no.5588, pp.746-748, 1976.

[12] S. Dupont and J. Luettin, Audio-visual speech modelling for continuous speech recognition, *IEEE Trans. on Multimedia*, vol.2, no.3, pp.141-151, 2000.

[13] J. N. Gowdy, A. Subramanya, C. Bartels and J. Bilmes, N-based muti-stream models for audio-visual speech recognition, *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.993-996, 2004.

[14] T.-L. Pao, W.-Y. Liao, Y.-T. Chen and T.-N. Wu, Mandarin audio-visual speech recognition with effects to the noise and emotion, *International Journal of Innovative Computing, Information and Control*, vol.6, no.2, pp.711-724, 2010.

[15] J. A. Bilmes and C. Bartels, Graphical model architectures for speech recognition, *IEEE Signal Processing Magazine*, vol.22, pp.89-100, 2005.

[16] J. L. Schwartz, F. Berthommier and C. Savariaux, Seeing to hear better: Evidence for early audio-visual interactions in speech identification, *ERIC Journal Articles: Reports-Research, Cognition*, vol.93, no.2, pp.69-78, 2004.

[17] J. Y. Kim, J. H. Lee and K. Shirai, An efficient lip-reading method robust to illumination variations, *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol.E85, no.9, pp.2164-2168, 2002.

[18] E. Saber and A. M. Tekalp, Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost function, *Pattern Recognition Letters*, vol.19, pp.669-680, 1998.

[19] N. Eyeno, A. Caplier and P. Y. Coulon, A new color transformation for lips segmention, *Proc. of IEEE International Conference on Acoustic, Speech, Signal Processing*, pp.557-560, 1993.

[20] R. Stiefelhagen, U. Meier and J. Yang, Real-time lip-tracking for lip reading, *Proc. of Eurospeech, the 5th European Conference on Speech Communication and Technology*, 1997.

[21] J. Yang, R. Stiefelhagne, U. Meier and A. Waibel, Real-time face and facial feature tracking and application, *Proc. of Auditory-Visual Speech Processing*, pp.79-84, 1998.

[22] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, An efficient k-means clustering algorithm: Analysis and implementation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.24, pp.881-892, 2002.

[23] G. Potamanos, H. P. Graf and E. Cosatto, An image transform approach for HMM based automatic lipreading, *Proc. of the International Conference on Image Processing*, Chicago, USA, vol.3, pp.173-177, 1998.

[24] C. C. Chibelushi, F. Deravi and J. S. Moson, A review of speech-based bimodal recognition, *IEEE Trans. on Multimedia*, vol.4, no.1, pp.23-37, 2002.

[25] P. Scanlon and R. Reilly, Feature analysis for automatic speechreading, *Proc. of the International Conference on Multimedia and Expo*, pp.625-630, 2001.

[26] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Pearson Prentice Hall, 1992.