International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.10 (2015), pp.337-346
http://dx.doi.org/10.14257/ijmue.2015.10.10.33

Feature Generations Analysis of Lip Image Streams for Isolate Words Recognition

Yong-Ki Kim[1], Jong Gwan Lim[2], Sahngwoon Lee[3] and Mi-Hye Kim[*4]
[1, 4] *Dept. of Computer Engineering, Chungbuk National University, CheongJu, Korea*
[2] *Dept. of Mechanical Engineering, KAIST, Daejeon, Korea*
[3] *Systran International*
{moodeath.kyk[1], jonggwanlim[2]}@gmail.com,
Sahngwoon,Lee@systrangroup.com[3], mhkim@chungbuk.ac.kr[4]

Abstract

To overcome the decrease in the recognition rate of voice recognition in noisy environments, the implementation of Audio Visual Speech Recognition (AVSR), which combines voice and lip information, has been attempted since the 1990s. This study aims to investigate the discrimination of various features extracted from lip image data using dynamic time warping (DTW) as an objective function to implement a robust lip-reading system as the core process of AVSR. The features taken from existing literature are grid-based features, including gray level, optical flow, and Sobel operator gradient, and various ratios of lip shapes calculated based on coordinates. According to the results of the application of DTW to respective feature generation methods using 180 pieces of data collected from ten study subjects who each uttered six isolated words three times, the mean recognition rate was found to be up to 60.55%. The feature that showed the highest recognition rate was the combined vector of a width/height ratio of the outer lip and the height of the inner lip, and grid-based features were found to outperform coordinate-based features in the recognition rate of certain words.

Keywords: *Lip-Reading System, AVSR, Image Processing, Isolated Words*

1. Introduction

Human beings generally use voice as a means of communicating with one another. In addition to that, as information and communication technology had advanced rapidly, voice as a means for interaction between humans and computers has emerged as an effective means of information exchange between humans and machines. Because a user interface with voice has advantages of being more intuitive and capable of various types of interactions than a user interface with a keyboard. Research on automatic speech recognition (ASR) has been actively conducted. However, ASR has the shortcoming of a decreased recognition rate in real-life environments where noise exists. To overcome the decrease in the recognition rate of voice recognition in noisy environments, Audio Visual Speech Recognition (AVSR), which combines voice and lip information, has been proposed. As it was found that speech recognition performance in a noisy environment is improved with lip reading rather than using only voice information, the lip-reading system has been an interesting study subject since the mid-1990s, and it has become recognized as one of the alternatives for improving voice recognition in a noisy environment [1-2, 10]. The basic processing for the recognition of a specific word through the lip-reading system starts by extracting feature points, which is largely divided into a method that uses the image information of all articulators and a method that uses only the lip information. Early studies have focused on spatiotemporal changes by taking the entire image value of

ISSN: 1975-0080 IJMUE

the lip area [1, 3], and then efforts to compress the amount of information by eliminating redundancy, such as through principal component analysis and discrete cosine transformation have been incorporated [2]. Approaches have been proposed, such as representing lip contours with an active shape model, active appearance model, snake-shape model, or convex hull, among others. By limiting key features during utterance to lip contours [4, 6, 8-10] or using gray level, Sobel, and optical flow as features while using the image information of all articulators, but specializing in contrast, edge, and lip motion [2, 5, 7]. In terms of readers, in most studies, the readers handling multidimensional time-series data, including the Hidden Markov Model (HMM), Support Vector Machine, and artificial neural network, has been commonly used.

This study aims to investigate the discrimination of various features extracted from lip image data prior to studying the audio-visual speech recognition (AVSR) system.

2. Features of Mouth Shape for the Korean Language

A syllable as a phonological unit in Korean is a combination of a vowel and a consonant, and the shape of the lips in the vocalization process is largely determined by the type of vowel rather than the consonant. Regarding vowels, the sound is determined by the degree of upper and lower-lip movement, position of the tongue, and exposure of the teeth, among other factors. In particular, the vowels that are differentiated by the degree of upper and lower-lip movement and the width of the mouth are shown in Figure 1 [12].

Table 1. Classification of Korean Consonants by Place of Articulation

Sounds	Consonants
labial sound	ㅂ/b/, ㅍ/p/, ㅁ/m/
lingual sound	ㄴ/n/, ㄷ/d/, ㅌ/t/, ㄹ/l,r/
dental sound	ㅅ/s,ʃ/, ㅈ/z/, ㅊ/tʃ/
velar sound	ㄱ/g/, ㅋ/k/
guttural sound	ㅇ/ŋ/, ㅎ/h/

On the other hand, regarding consonants, other than labial sounds, which are the sounds made with closed lips (upper and lower lips), such as ㅁ/m/, ㅂ/b/, and ㅍ/p/ in Table 1, no visual features of the lip shapes are observed [11-12]. This is different from the commonly restricted role of consonants in ASR.

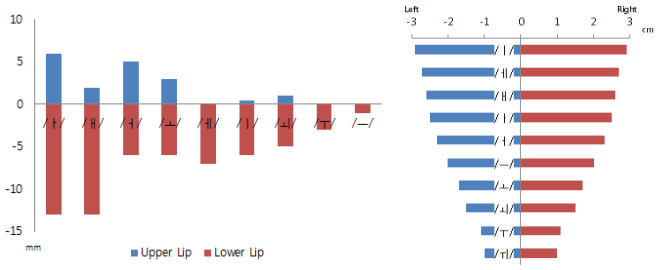

Figure 1. Lip Movement by Vowel. Height Difference between Upper and Lower Lips by Vowel (left), and Difference in Width of Lips by Vowel (right)

3. Generation of Various Mouth-Shape Features

Figure 2 is a flowchart of the entire process of mouth-shape recognition conducted in this study.

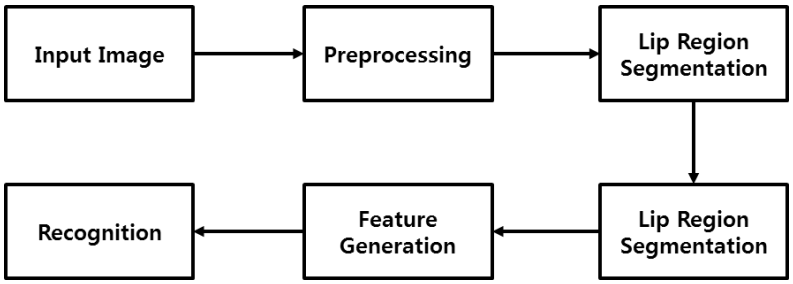

Figure 2. Flowchart of Entire Process of Mouth-Shape Recognition

To generate the feature points, the face area, eye area, and lip area are detected, and then the utterance zone is detected. Before the detection of the face area, the RGB color space is converted to YCbCr color space, the brightness of which is then corrected with histogram equalization. To detect the lip area, the face area is detected using an elliptic skin model and AdaBoost [13], and then the eye area is detected within the detected face area. In the eye area, a dark area is detected by binarizing the difference between the close-operated image and the original image in the gray-level image [14]. Based on the detected eye area, the lip area is detected using the geometric model of the face (Figure 3) [15].

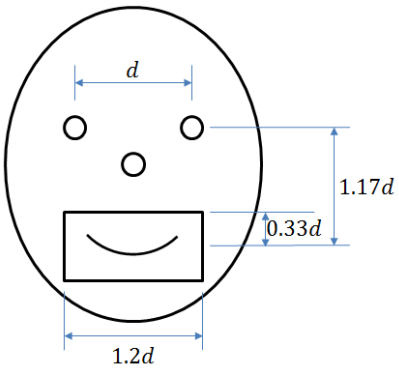

Figure 3. Geometric Model of Face

To detect the utterance zone, the lip contours are approximated, and then, among specific coordinates, the mean of the three features with the highest variances of coordinates per frame during utterance is obtained (Figure 4). At this time, considering the phenomenon in which the lips are closed briefly during utterance due to the pronunciations containing the consonants of labial sounds, signal smoothing is added. The utterance zone is then detected by applying the threshold [16].

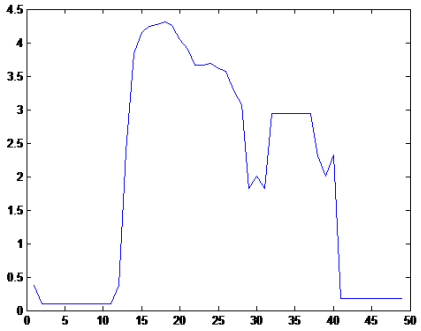

Figure 4. Mean of Lip-Shape Coordinate Features

In this study, mouth-shape features are generated separately for grid-based feature vectors that reflect the overall movement of articulators, such as the lips and tongue, and coordinate-based feature vectors that express only the movement of the lips. First, regarding grid-based feature generation, the same-sized grids are overlaid on the lip area, as shown in Figure 4, and then the gray level, optical flow, and Sobel operator gradient of the pixel value of each grid are selected.

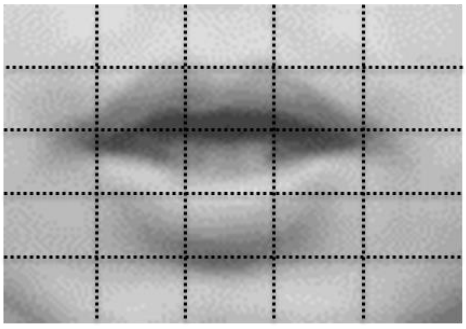

Figure 5. Lip Area Overlaid with Grid

The gray-level feature vectors are used to classify the contrast of articulators inside the lips that change during utterance [1]. Optical flow is proposed to represent the movement of the lips and the area around them during utterance, and the Lucas–Kanade [17] method, which is widely used for optical flow estimation, is used. The Sobel operator gradient used to extract the edge of an object in an image extracts the boundaries of all articulators [18].

The gradient feature vectors with the gray level, optical flow, and Sobel operator are normalized into the means for each feature within a grid, as shown in Eq. (4)–(6), to compensate for the effect of lighting.

$$F_{gray} = [\overline{g_1}\ \overline{g_2}\ \ldots\ \overline{g_n}]^T \tag{1}$$

$$F_{opticalflow} = [\overline{o_1}\ \overline{o_2}\ \ldots\ \overline{o_n}]^T \tag{2}$$

$$F_{sobel} = [\overline{s_1}\ \overline{s_2}\ \ldots\ \overline{s_n}]^T \tag{3}$$

$$\overline{g_n} = \frac{\sum_{(x,y)\in C_n} g(x,y)}{C_{pixel}} \tag{4}$$

$$\overline{o_n} = \frac{\sum_{(x,y)\in C_n} o(x,y)}{C_{pixel}} \tag{5}$$

$$\overline{s_n} = \frac{\sum_{(x,y)\in C_n} s(x,y)}{C_{pixel}} \tag{6}$$

Here, $\overline{g}$, $\overline{o}$, and $\overline{s}$ denote the means of the sizes of the gray level, optical flow, and Sobel operator gradient in each grid, respectively; $g(x,y)$, $o(x,y)$, and $s(x,y)$ denote the sizes of the gray level, optical flow, and Sobel operator gradient in pixel units in the grid, respectively; and C_n denotes the set with pixels in the grid as elements. In addition, C_{pixel} denotes the number of pixels inside the grid.

The coordinate-based feature vectors are generated by manually detecting the coordinates of 16 points that approximate the lip contours based on certain rules. As shown in Figure 5, the coordinate detection criteria include identifying the end points at both lateral sides (points 0 and 4), and then identifying the points at which the vertical line from the center of the straight line that connects the two lateral end points meets the outer edge of the lips (points 2 and 6). From these four points, the other four points are placed at regular intervals. The coordinates of the inner lips are placed in the same manner, starting from the later end points of the inner lips (points 8 and 12).

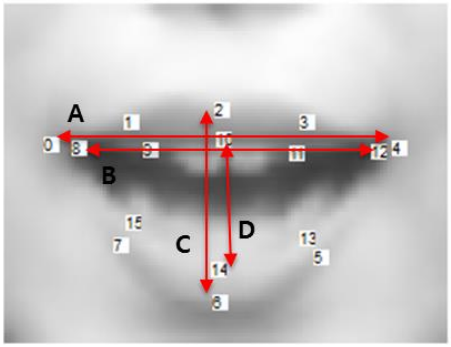

Figure 6. Coordinate-Based Features (A: Outer-Lip Width, B: Inner-Lip Width, C: Outer-Lip Height, D: Inner-Lip Height)

Using the lip-shape coordinates detected using the above method, the feature vectors are comprised of the width/height ratio of the outer lip (p_1), the width/height ratio of the inner lip (p_2), the width of the outer lip (p_3), the height of the outer lip (p_4), the width of the inner lip (p_5), and the height of the inner lip (p_6), as shown in Eq. (7).

$$F_{point} = [p_1 \ p_2 \ p_3 \ p_4 \ p_5 \ p_6]^T \tag{7}$$

$$p_1 = \frac{\frac{C^i}{C^0}}{\frac{A^i}{A^0}} \tag{8}$$

$$p_2 = \frac{\frac{D^i}{D^0}}{\frac{B^i}{B^0}} \tag{9}$$

$$p_3 = \frac{A^i}{A^0} \tag{10}$$

$$p_4 = \frac{C^i}{C^0} \tag{11}$$

$$p_5 = \frac{B^i}{B_i^0} \tag{12}$$

$$p_6 = \frac{D^i}{D^0} \tag{13}$$

Here, i denotes the frame of the image, and, accordingly, 0 is the respective length of the lip detected in the first frame. A denotes the width of the outer lip, C denotes the height of the outer lip, B denotes the width of the inner lip, and D denotes the height of the inner lip.

4. Experiment and Results

A total of 10 subjects (5 men and 5 women) were recruited for the study. The isolated words used in the experiment consisted of 하이갤럭시 /haigælʌksi/ (w1), 하이스마트폰 /haismɑ:rtpon/ (w2), 하이카메라 /haikamera/ (w3), 하이메시지 /haimesɪdʒ/ (w4), 하이카카오톡 /haikakaotok/ (w5), and 하이전화걸기 /haiʒʌnhwagʌlgi/ (w6). The 10 participants uttered each word three times, resulting in a total of 180 pieces of collected data. The distance between the speaker and the camera was 30–50cm, and the speaker's face and the camera were fixed on an imaginary straight line. Dynamic time warping (DTW) was used to determine the discrimination of various features extracted from the previously discussed lip image data. DTW is the representative method used to obtain the similarities between the reference time-series pattern and the time series with input features of different lengths, and it is a suitable method for comparing time-series data. Conventionally, the Hidden Markov Model (HMM) is used, because HMM requires data over a certain size for learning, and DTW was chosen for the present study with small data [19-20].

First, to determine the optimal size of the grid in the grid-based method, the recognition rates for different grid sizes were compared (Figure 6). Based on the results, the grid of the 5 x 5 size showing the highest mean recognition rate was selected.

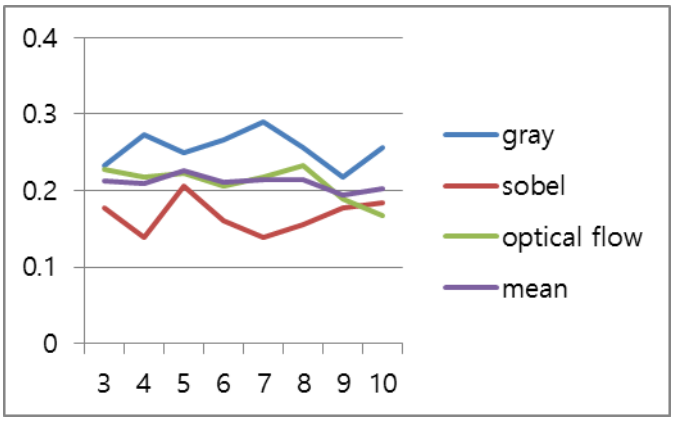

Figure 7. Changes in Recognition Rate of Grid-Based Features by Grid Size

To find the optimal feature combination for coordinate-based features, a recognition experiment was conducted by creating various vector combinations by changing the number of dimensions from one to six using the variables shown in Eqs. (7) to (13). As shown in Figure 8, the results showed that the mean and the lowest recognition rates tended to increase, but the highest recognition rates decreased as the number of dimensions increased.

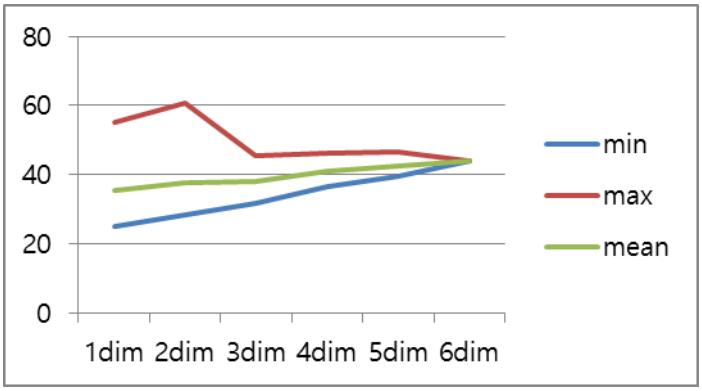

Figure 8. Changes in Recognition Rate by Number of Dimensions

Table 2 shows the feature vector combinations that showed the highest and lowest recognition rates for each number of dimensions. In the experiment, the width/height ratio of the outer lip and the height of the inner lip were effective features, and then a comparative recognition experiment was conducted using the $\{p_1, p_6\}$ vector combination as the feature.

Table 2. Highest and Lowest Feature Vectors by Number of Dimensions

Dim	1	2	3	4	5	6
Highest	$\{p_6\}$	$\{p_1, p_6\}$	$\{p_1, p_4, p_6\}$	$\{p_2, p_4, p_5, p_6\}$	$\{p_1, p_2, p_4, p_5, p_6\}$	$\{p_1, p_2, p_3, p_4, p_5, p_6\}$
Lowest	$\{p_3\}$	$\{p_1, p_2\}$	$\{p_1, p_3, p_5\}$	$\{p_1, p_3, p_4, p_5\}$	$\{p_1, p_2, p_3, p_4, p_5\}$	$\{p_1, p_2, p_3, p_4, p_5, p_6\}$

Finally, the results of the recognition experiment for each feature generation method are shown in Figure 9. On average, the coordinate-based method recorded the highest recognition rate, but the grid-based method showed superior recognition rates for certain

words. Specifically, the Sobel operator gradient feature and optical flow feature recorded higher recognition rates for w1 and w4, and the gray-level feature recorded a higher recognition rate for w4 than the coordinate-based method.

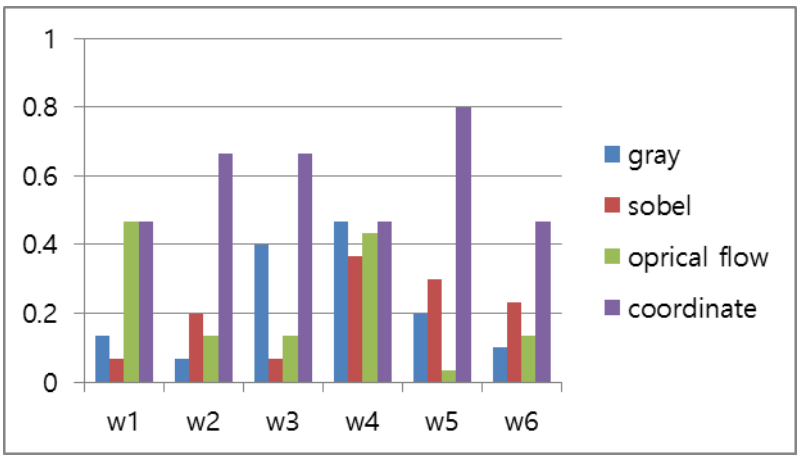

Figure 9. Mouth-Shape Recognition Results by Features

Table 3. Mouth-Shape Recognition Confusion Matrix

	W1	W2	W3	W4	W5	W6	Recognition Rate
W1	14	5	2	2	4	3	46.7%
W2	4	21	2	1	2	0	70%
W3	0	1	20	7	0	2	66.7%
W4	4	5	5	15	0	1	50%
W5	2	2	2	0	24	0	80%
W6	4	6	0	3	2	15	50%

Table 3 shows the confusion matrix for the coordinate-based feature that recorded the highest recognition rate among various features. The recognition results showed high recognition rates for class w2 and class w5. It was concluded that class w2 showed strong coordinate-based features, because it had more labial articulation than other words in the experiment, containing ㅁ/m/ and ㅍ/p/, and class w5 showed a high recognition rate as it consisted of only the vowels ㅗ/o/ and ㅏ/a/, which were not included in the other classes. It was concluded that the reason the grid-based features represented by gray level outperformed the coordinate-based features in recognition rate is because when the consonant, ㄹ/l/, met the vowel, ㅏ/a/, a noticeable change in the movement of the tongue was observed in the relatively widely opened mouth. As this is a new observation in which the utterance of a consonant appears to influence the recognition rate, additional verification is needed.

5. Conclusions

In the AVSR system, the performance of the lip-reading system is the key component that determines the entire system. The comparison of isolated word recognition rates through lip reading shows that higher average recognition rates were obtained by using only the information on lip-contour movements rather than comprehensive information,

such as that on lips and the tongue. On average, the coordinate-based feature with the highest recognition rate was found to be the vector combination of the width/height ratio of the outer lip and the height of the inner lip. However, the feature using gray level shows higher recognition rates than the coordinate-based feature for certain isolated words, and the feature point using optical flow showed a similar recognition rate to the coordinate-based feature point. It was also found that labial articulation during utterance can be a major feature in word recognition, unlike voice recognition, which is mainly affected by vowels. However, additional research is needed concerning which consonants or vowels relate to the position of the tongue or overall movement of the lips as key features.

Finally, this study shows the potential for a vowel classification system that uses only image data, and it is expected to make a significant contribution to the performance improvement of voice recognition in noisy environments. Resolving the issues described earlier and conducting additional studies are expected to lead to the building of a more robust AVSR system.

References

[1] Chien S. I. and Bae K. S., "A study on the performance improvement of lip reading and speech recognition by fusing audio/visual information", Technical Report, (in Korean), **(2002)**.
[2] Min G. S., "A Study on the Robust Lip Detection and Lip Reading System for Audio-Visual Speech Recognition in Mobile Environment", Graduate School of Chonnam National University doctoral dissertation. (in Korean).
[3] Lucey S., Sridharan S. and Chandran V., "Adaptive mouth segmentation using chromatic features", Pattern Recognition Letters, vol. 23, no.11, **(2002)**, pp.1293–1302.
[4] Wang S. L., Lau W. H., Leung S. H. and Yan, H.: "A real-time automatic lip reading system", Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on, IEEE, vol. 2 vol. 2, **(2004)**.
[5] Lay Y. L., Tsai C. H., Yang H. J., Lin C. S. and Lai C. Z., "The application of extension neuro-network on computer-assisted lip-reading recognition for hearing impaired", Expert Systems with Applications, vol. 34, no. 2, **(2008)**, pp. 1465–1473.
[6] Bagai A., Gandhi H., Goyal R., Kohli M. and Prasad T. V., "Lip-reading using neural networks", International Journal of Computer Science and Network vol. 9, no.4, **(2009)**, pp.108–111.
[7] Shaikh A. A., Kumar D. K., Yau W. C., Che A. M. Z. and Gubbi J., "Lip reading using optical flow and support vector machines", In Image and Signal Processing (CISP), 2010 3rd International Congress on, IEEE, vol. 1, **(2010)**, pp. 327–330.
[8] Skodras E. and Fakotakis N., "An unconstrained method for lip detection in color images", In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, **(2011)**, pp.1013–1016.
[9] Ibrahim M. Z. and Mulvaney D. J., "Robust geometrical-based lip-reading using Hidden Markov models", In EUROCON, 2013 IEEE, **(2013)**, pp.2011–2016.
[10] Dave M. N. and Patel N. M., "Phoneme and Viseme based Approach for Lip Synchronization", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 7, no. 3, **(2014)**, pp.385–394.
[11] "The National Institute of the Korean Language", http://www.korean.go.kr/front_eng/main.do
[12] Kim Y. K., Lim J. G. and Kim M. H., "Lip Reading Algorithm Using Bool Matrix and SVM", International Conference on Small & Medium Business, (in Korean), **(2015)**, pp.267–268.
[13] Man L., Xiao-yu W. and Hui-ling M., "Face Automatic Detection based on Elliptic Skin Model and Improved Adaboost Algorithm", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, no. 2, **(2015)**, pp.227–234.
[14] Petajan E., "Automatic lip reading to enhance speech recognition", Illinois at Urbana-Champaign University Doctoral dissertation, **(1984)**.
[15] Chien S. I. and Choi I., "Face and facial landmarks location based on log-polar mapping", Biologically Motivated Computer Vision. Springer Berlin Heidelberg, **(2000)**, pp.379–386.
[16] Lim J. G., Kim M. H. and Lee S., "Empirical Validation of Objective Functions in Feature Selection Based on Acceleration Motion Segmentation Data", Mathematical Problems in Engineering, 501, 280140, **(2015)**.
[17] Sun D., Roth S. and Black M. J., "Secrets of optical flow estimation and their principles", Computer Vision and Pattern Recognition (CVPR) 2010 IEEE Conference on, **(2010)**, pp.2432–2439.
[18] Kanopoulos N., Vasanthavada N. and Baker R. L Design of an image edge detection filter using the Sobel operator", Solid-State Circuits, IEEE Journal of, vol. 23, no. 2, **(1988)**, pp.358–367.

[19] Lim J. G., Sohn Y. I. and Kwon D. S., "Real-time Accelerometer Signal Processing of End Point Detection and Feature Extraction for Motion Detection", In Analysis, Design, and Evaluation of Human-Machine Systems, 2007, 10th IFAC/ IFIP/ IFORS/ IEA Symposium on. IFAC, **(2007)**.
[20] Rabiner L. R. and Juang B., "Fundamentals of Speech Recognition", Prentice-Hall, Inc., (Chapter 4), **(1993)**.

