Journal of the Korea Industrial Information Systems Research Vol.25 No.4, Aug. 2020 :9-16
http://dx.doi.org/10.9723/jksiis.2020.25.4.009 ISSN:1229-3741

베이지안 분류를 이용한 립 리딩 시스템+

(Lip-reading System based on Bayesian Classifier)

김 성 우[1], 차 경 애[2]*, 박 세 현[3]
(Seong-Woo Kim, Kyung-Ae Cha, and Se-Hyun Park)

요 약 음성 정보를 배제하고 영상 정보만을 이용한 발음 인식 시스템은 다양한 맞춤형 서비스에 적용될 수 있다. 본 논문에서는 베이지안 분류기를 기반으로 입술 모양을 인식하여 한글 모음을 구분하는 시스템을 개발한다. 얼굴 이미지의 입술 모양에서 특징 벡터를 추출하고 설계된 기계 학습 모델을 적용하여 실험한 결과 'ㅏ' 발음의 경우 94%의 인식률을 보였으며, 평균 인식률은 약 84%를 나타내었다. 또한 비교군으로 실험한 CNN 환경에서의 인식률보다 높은 결과를 보였다. 이를 통해서 입술 영역의 랜드 마크로 설계된 특징 값을 사용하는 베이지안 분류 기법이 적은 수의 훈련 데이터에서 보다 효율적일 수 있음을 알 수 있다. 따라서 모바일 디바이스와 같은 제한적 하드웨어에서 응용 가능한 어플리케이션 개발에 활용할 수 있다.

핵심주제어: 베이지안 분류, 독순술, 모음 인식, 얼굴 랜드마크, 기계 학습

Abstract Pronunciation recognition systems that use only video information and ignore voice information can be applied to various customized services. In this paper, we develop a system that applies a Bayesian classifier to distinguish Korean vowels via lip shapes in images. We extract feature vectors from the lip shapes of facial images and apply them to the designed machine learning model. Our experiments show that the system's recognition rate is 94% for the pronunciation of 'A', and the system's average recognition rate is approximately 84%, which is higher than that of the CNN tested for comparison. Our results show that our Bayesian classification method with feature values from lip region landmarks is efficient on a small training set. Therefore, it can be used for application development on limited hardware such as mobile devices.

Keywords: Bayesian classifier, Lip-reading, Recognition of vowel pronunciation, Facial landmark, Machine learning

* Corresponding Author: chaka@daegu.ac.kr
+ This research was supported by the Daegu University Research Grant, 2018.
Manuscript received June 15, 2020 / revised July 30, 2020 / accepted August 03, 2020
1) Department of Computer & Communication Engineering, Daegu University, 1st Author
2) School of ICT Convergence, Daegu University, Corresponding Author
3) School of ICT Convergence, Daegu University, 3rd Author

1. Introduction

Recognizing human pronunciation in images can be useful for various personalization services. Researchers have detected human faces in video and extracted desired facial areas to distinguish facial expressions and pronunciations. The main technique for

recognizing pronunciation based on facial imagery is mouth-shaped detection (Gyu et al., 2009; Xianoyi, 2017).

Mouth-shaped detection typically involves recognizing words or pronunciations by extracting mouth shapes from images and analyzing them together with sound information (Lee et al., 2002; Çetingül et al., 2006; Lim et al., 2018). However, combining mouth shape and voice information is disadvantageous because the recognition rate varies depending on the quality of the voice information, which is heavily influenced by noise.

In this paper, we distinguish five Korean pronunciations based on human lip shape in images without voice information. We identify pronunciations with a Bayesian classifier by detecting the mouth area in the human face and then using the distance between the upper lip and lower lip, around the jaw and mouth, and so on. To obtain the facial objects, we use the Haar Cascade algorithm to detect the location of the eyes, nose and mouth and measure the difference in position between the lips or between the mouth and the jaw by assigning feature points to the detected facial area. Finally, these distances are used as attributes to learn from a Bayesian classifier, and the relevant pronunciations are distinguished.

2. Related Work

Chung et al. (2016) recognized the words being spoken by a talking face given video but not audio information. They described an automatic pipeline system for collecting and processing a visual speech recognition dataset from television programs and reported excellent lip-reading results using a CNN (Convolutional Neural Network) architecture.

Other studies have used algorithms such as deep learning for detecting feature vectors of the mouth region uttering experimental words. (Kim, 2016; Xianoyi, 2017; Kim, 2018). Similar studies include mouth shape recognition and pronunciation discrimination using SVMs (Support Vector Machines) and Boolean Matrices (Kim et al., 2014).

The Cascade method has the advantage of quickly detecting readily detectable characteristics with a simple detector, which can improve the speed of detecting objects in an image. The Haar Cascade (Viola and Jones, 2001), which detects features by using the difference in brightness between regions of interest in an image, is frequently used to recognize human faces and detect objects such as the eyes, nose, and mouth (Hwang, 2017).

In this paper, we design an effective feature vector for recognizing Korean vowels and apply it to a Bayesian learning model (Choi et al., 2001; Oh, 2008). The experimental results show the advantage of our proposed method.

3. Implementation of the Vowel Pronunciation Recognition System

Fig. 1 shows the processes of our system. Based on a designed ML (Machine Learning) model (Kim et al., 2019), this paper presents an implementation technique of the proposed system along with an experimental analysis to present and verify the extracted data.

The facial landmarks defined in *dlib* (Dlib, 2002) are applied to the frontal face objects detected by the Haar Cascade, and the distances of the landmarks on the face objects are the random vector of the Bayesian classifier.

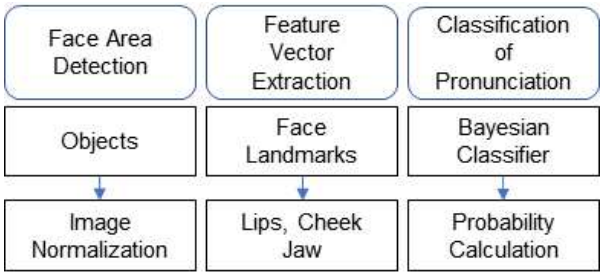

Fig. 1 System Structure Diagram

As shown in Fig. 2, the face region is first converted into four coordinate points as a rectangle, representing a ROI (Region Of Interest). After normalizing the ROI image to a constant size, landmarks are given to the facial area by loading the pre-trained data in the *dlib* library. The landmarks from 48 to 68 points represent the lip shape of a vowel pronunciation.

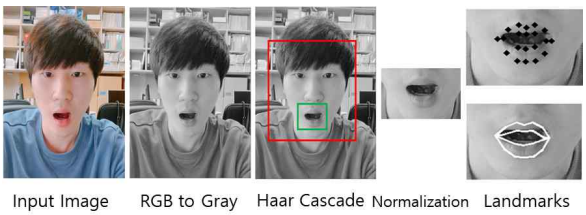

Fig. 2 Lip Shape Detection

The feature attributes used in the Bayesian classification are defined as follows: 'x1'-horizontal length of mouth, 'x2'-vertical length of mouth, 'x3'-thickness of upper lip, 'x4'-thickness of lower lip, 'x5'-horizontal length of chin, and 'x6'-vertical length of chin. Fig. 3 illustrates the visual signification of each attribute.

Naive Bayesian theory classifies random values into specific classes on the premise that the properties are independent of each other. The effect of the corresponding attribute value on each class is measured to enable the classification of random values.

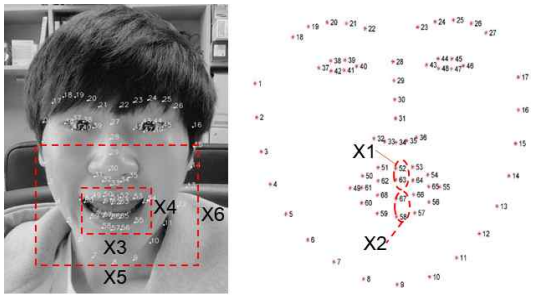

Fig. 3 Feature Attribute Definition from Face Object Landmark

The pronunciation of 'A', 'I', 'U', 'E' and 'O' can be defined in each class, and feature vectors extracted from human face images can be classified into one of the five classes.

In this paper, we define the vowel pronunciation class as $C_0 = [A], C_1 = [I], C_2 = [U], C_3 = [E], C_4 = [O]$.

The random input values to the Bayesian classifier, $X = (X_k | 0 \le k \le n), n = 5$, are defined by the six attributes characterizing the mouth shape of the corresponding pronunciation.

The prior probability density function, $P(C_i)$ is the probability distribution of a class that determines the vowel with a random feature vector, X belongs in. It is assumed to follow a Gaussian distribution. When μ_i is a mean value, and $\sum_i$ is a covariance value of its corresponding class, C_i, the likelihood can be expressed by the following equation (1).

$$P(X \mid C_i) = P(x_0, \ldots x_5 \mid C_i) = \prod_{k=0}^{5} P(x_k \mid C_i) \quad (1)$$

$$P(x_k \mid C_i) \propto N(x_k, \mu_i, \sum_i)$$

$$= \frac{1}{(2\pi)^{d/2} \left| \sum_i \right|^{1/2}} exp(-\frac{1}{2}(x_k - \mu_i)^T \sum^{-1} (x_k - \mu_i))$$

The derivation of a Gaussian distribution using actual feature vector data can be seen in Kim et al. (2019).

4. Development and Experimental Results

The proposed system was implemented with Python and OpenCV on a Windows 10 platform.

The input image set is a collection of 120 images for each of the 5 pronunciations. A total of 600 data images of men and women between the ages of 20 and 50 were used for the input data. Fig. 4 shows several images used in the experiments.

To apply the facial landmarks with virtually 100% accuracy, the faces were recognized within an angle of approximately 10 degrees when viewed from the front of the camera. Faces were always recognized from the boundary of the hair to the chin. Thus, the landmark extraction process shows an accurate value for each image. Since the ROI for the human face is accurately recognized, it is not affected by the background image. However, the lighting conditions were maintained such that the human faces were clearly visible.

Fig. 4 Sample Input Images

The distance of each feature attribute determined by the landmarks is measured in pixels. The set of generated feature tuples is trained by applying them as attribute values to the Bayesian classifier. When new data arrives, the corresponding pronunciation can be estimated as a Naive Bayes-based probability model.

For example, Table 1 shows the average of the pixel values of the measured feature attributes corresponding to the 5 pronunciations in the sample images.

We obtained statistical values for verifying the values of 500 of 600 images with measured feature vectors to determine the probability functions of each pronunciation.

Table 1 Measurement Values of Vowels' Pronunciation (unit : px)

X	A	I	U	E	O
x1	84.38	99.79	75.01	95.26	74.33
x2	97.98	113.85	101.03	97.07	106.54
x3	25.64	23.38	25.75	22.92	28.32
x4	85.04	48.19	52.07	49.38	59.74
x5	143.86	143.02	126.05	126.19	127.43
x6	180.86	202.37	198.0	190.91	204.1

Table 2 shows the mean and standard deviation values of attribute x1 using 500 tuples of measured pixel values.

Table 2 Statistical Values of x1 Attribute for Each Vowels' pronunciation (unit : px)

Vowel	Mean	Standard Deviation
A	90.799	10.68759
I	109.765	10.18144
U	79.01087	7.313586
E	104.5566	6.46112
O	77.34125	11.05134

The attribute x1 values indicate that it is particularly easy to distinguish between 'A' or 'U' pronunciation and 'E' pronunciation.

We can examine the statistical values for every feature attribute from x1 (see Table 1) to x6. With respect to the 'I' pronunciation, the value of x5 is relatively large because the mouth opens horizontally and the cheeks widen. The major difference between the similar mouth shapes for 'U' and 'O' is the length between the mouth and chin and the length between the cheeks. When changing from 'U' to 'O' pronunciation, the horizontal

length of the chin, x5, decreases, and the vertical length to the mouth and chin, x6, increases greatly.

We can also examine whether the recognition rate increases with the number of training images. The training data are composed of 10, 20, 40, and 60 images per pronunciation, so there are 50, 100, 200, and 300 images for each vowel, respectively. Moreover, The test set consists of 50 images for each pronunciation.

As shown in Table 3, there are many similar attribute values for each pronunciation, and even the same pronunciation varies greatly in each attribute value according to the speaker; thus, more accurate classification is possible as the number of training cases increases.

Table 3 Recognition Rate according to Number of Training Images

Training set (number)	50	100	200	300
Recognition Ratio(%)	49	68	78	82

Table 4 Confusion Matrix of Recognition Results

		Recognized class				
		A	I	U	E	O
Vowel	A	24				1
	I		24		4	
	U			11		5
	E		5		24	1
	O		1	4		19

To find out the ratio of incorrectly recognized pronunciations, 123 images were randomly selected from the 600 images. The recognition results are represented in Table 4. For example, there are 5 cases where 'U' is mistaken for 'O'. In this experiment, the 'O' and 'U' pronunciations are often mistaken for each other.

Fig. 5 shows the results of extracting each pronunciation from video frames. This is a smartphone application implementing the proposed system using smartphone camera video input. It is intended for experiments in real time vowel recognition because video data enable a continuous learning phase. In previous experiments, it took 30 seconds for the recognition result to reach 83% with detecting an ROI. Reflecting upon this result, we used approximately 30 seconds of video to recognize 5 different pronunciation.

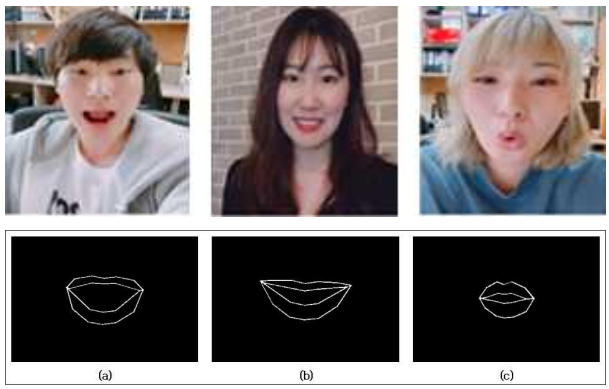

Fig. 5 Recognition Results from a Video
(a) 'A', (b) 'E', (c) 'O'

Finally, we compared a CNN method with our proposed system.

The input data of the CNN is an image generated by connecting points of landmarks corresponding to the mouth region.

Fig. 6 shows the processing steps of input image organization for the CNN network. Each point from 48 to 68 of the landmarks representing lip shape is connected by a white straight line. Then, the background area, except for the white lines in the lip image, is changed to black and used as a CNN input image. Since the input image is a gray level image, only one channel is implemented. For a set of 600 images, 500 images were divided into training data, and the remaining 100

images were used as test data.

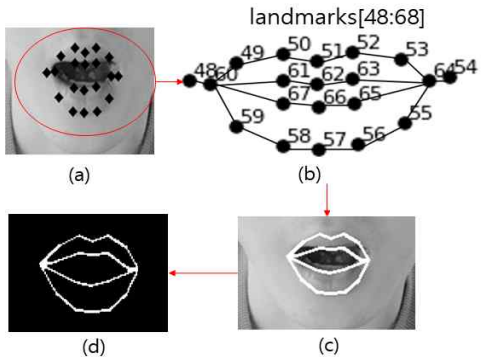

Fig. 6 Lip Shape Image Generation Processes
(a) Landmarks on the lip area
(b) Detail of landmarks [48:68]
(c) A captured image connected the landmarks with white lines
(d) Image with background changed to black

The classification network has the structure shown in Fig. 7. Instead of using a typical network such as LeNet or ResNet, we use a lightweight layers by adjusting parameters.

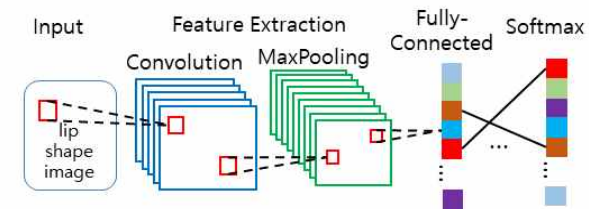

Fig. 7 CNN Layer

Other used techniques for activation and overfitting avoidance are shown in Table 5.

Table 5 Applied Functions to CNN

Function	Applied function
Activation	ReLU
Overfitting avoidance	Dropout
Optimizer	Adadelta

As shown in Table 6, for the proposed system, the 'A' pronunciation achieves the highest recognition rate at 94%, and for CNN, the 'U' pronunciation achieves the highest recognition rate at 82%.

Table 6 Comparison of CNN and Proposed Bayesian Classifier

Vowel	A	I	U	E	O
Bayesian	94%	87%	76%	80%	81%
CNN	50%	50%	82%	41%	30%

In all the vowel recognition results except 'U', the proposed system outperforms the CNN. The CNN system requires more than 5000 training data instances, so its recognition rate may have decreased due to the small training data set used.

Compared to the deep learning technique, the proposed method carries the advantages of efficient training time and non-reliance on high-spec hardware. Therefore, the proposed system is effective in an application field that can increase the pronunciation recognition rate with only a small number of learning images.

5. Conclusions

In this paper, a system was implemented to detect a human face in real-time images and distinguish five Korean vowel pronunciations according to lip shape.

Based on the experimental results, this study provides insight on considering a classic ML technique may perform better than deep learning method on a small sample dataset.

Moreover, through the implementation results, the pronunciation of Korean vowels can be recognized in video-only input data. This technology can be used in educational systems with lip-reading for the hearing

impaired. The method of learning to read lip shapes in the case of vowel pronunciations, can be provided by a video input interface, so as to practice and learn the accurate lip shapes of the pronunciations.

In addition, this technology can be utilized for custom input interfaces by learning the shape of an individual's pronunciation using classic ML techniques for limited hardware specifications such as smartphones.

References

Choi, J. H., Kim, J. B., Kim, D. G., and Rim, K. W. (2001). Bayesian Model for Probabilistic Unsupervised Learning, *Journal of Fuzzy Logic and Intelligent Systems*, 11(9), 849-854.

Çetingül, H. E., Erzin, E., Yemez, Y., and Tekalp, A. M. (2006). Multimodal Speaker/Speech Recognition using Lip Motion, Lip Texture and Audio, *Signal Processing*, 86(12), 3549-3558.

Chung, J. S., and Zisserman, A. (2016). Lip Reading in the Wild, *Asian Conference on Computer Vision*, Springer, Cham.

Dlib C++ Library (2002). *General Purpose Cross-platform Software Library,* http://dlib.net/ (Accessed on Aug. 10th, 2020).

Gyu, S. M., Pham, T. T., Kim, J. Y., and Taek, H. S. (2009). A Study on Lip Detection based on Eye Localization for Visual Speech Recognition in Mobile Environment, *International Journal of Fuzzy Logic and Intelligent Systems*, 19(4), 478-484.

Hwang, W. (2017). Research Trends in Deep Learning Based Face Detection, Landmark Detection and Face Recognition, *Broadcasting and Media Magazine*, 22(4), 41-49.

Kim, Y. K., Lee, H. S., and Kim, M. H. (2014). Lip Reading Method using Bool Matrix and SVM, *Proceedings of 2014 Conference on Korea, HCI*, pp. 179-182.

Kim, Y. K., Lim, J. G., and Kim, M. H. (2016). Lip Reading Method using CNN for Utterance Period Detection, *Journal of Digital Convergence*, 14(8), 233-243.

Kim, D., Choi, S., and Kwak, S. (2018), Deep Learning Based Fake Face Detection, *Journal of the Korea Industrial Information Systems Research*, 23(5), 9-17.

Kim, S., Cha, K., and Park, S. (2019). Recognition of Korean Vowels using Bayesian Classification with Mouth Shape, *Journal of Korea Multimedia Society*, 22(8), 852-859.

Lee, S., Lee, Y., Hong, H., Yun, B., and Han, M. (2002), Audio-visual Integration based Multi-modal Speech Recognition System, *Proceedings of KIPS Fall Conference*, 707-710.

Lim, D. Y., Kim, S. G., and Chong, K. T. (2018). Development of a Real-time Lip Recognition for Improving English Pronunciation using Deep Learning, *Journal of Institute of Control, Robotics and Systems*, 24(4), 327-333.

Oh, I. S. (2008). *Pattern Recognition*, Kyobobook.

Viola, P., and Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features, *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001(1), 511-518.

Xianoyi, Y. (2017). *Lipreading Recognition of English Vowels using Convolutional Neural Network and Recurrent Neural Network*, Master's Thesis, Chonbuk National University, Korea.

김 성 우 (Seong-Woo Kim)

- 대구대학교 정보통신대학 정보통신공학부 학사
- 대구대학교 정보통신공학과 석사
- 관심분야: 영상처리, 스마트어플리케이션, 인공지능, 딥러닝

차 경 애 (Kyung-Ae Cha)

- 종신회원
- 경북대학교 컴퓨터과학과 학사
- 경북대학교 컴퓨터과학과 석사
- 경북대학교 컴퓨터과학과 박사
- (현재) 대구대학교 정보통신대학 ICT융합학부 교수
- 관심분야: 멀티미디어시스템, 인공지능

박 세 현 (Se-Hyun Park)

- 종신회원
- 경북대학교 컴퓨터공학과 학사
- 경북대학교 컴퓨터공학과 석사
- 경북대학교 컴퓨터공학과 박사
- (현재) 대구대학교 정보통신대학 ICT융합학부 교수
- 관심분야: 컴퓨터비전, 인공지능