Advanced Science and Technology Letters
Vol.107 (Ubiquitous Science and Engineering 2015), pp.57-61
http://dx.doi.org/10.14257/astl.2015.107.14

Comparison of Lip Image Feature Extraction Methods for Improvement of Isolated Word Recognition Rate

Yong-Ki Kim[1], Jong Gwan Lim[2], Mi-Hye Kim*[3]

[1, 3] Dept. of Computer Engineering, Chungbuk National University, CheongJu, Korea
[2] Dept. of Mechanical Engineering, KAIST, Daejeon, Korea
{moodeath.kyk[1], jonggwanlim[2]}@gmail.com, mhkim@chungbuk.ac.kr[3]

Abstract. This study aims to investigate the discrimination of various features extracted from lip image data. A total of 90 pieces of data were collected as five subjects uttered six isolated words three times. The results of speech recognition through two different feature generation methods showed mean recognition rates of up to 60%. Although the grid-based feature extraction method yielded higher recognition rates for certain isolated words, the highest recognition rate was found to be the coordinate-based feature of the combined vector of width/height ratio of the outer lip and the height of inner lip.

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

1 Introduction

As it was found that speech recognition performance in a noisy environment is improved by combining voice information with mouth shape information rather than using only the 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 [1–10]. This study aims to investigate the discrimination of various features extracted from lip image data prior to studying the AVSR (Audio-Visual Speech Recognition) system.

2 Features of Mouth Shape for 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 consonant. Regarding consonants, other than the labial sounds, which are the sounds made with two closed lips (upper and lower lips), such as /ㅁ/, /ㅂ/, /ㅃ/, and /ㅍ/ (analogous to /m/, /b/, /pp/, and /p/), no visual features are observed. Regarding vowels, the sound is determined by the degree of upper lip and lower lip movement, the width of the mouth opening, and the height and position of the tongue [11].

ISSN: 2287-1233 ASTL

3 Generation of Various Mouth-Shape Features

Prior to the generation of feature points, facial area detection, eye area detection, and lip area detection are performed, which are followed by utterance period detection. For lip area detection, the facial area is detected using Adaboost, and then the eye area is detected based on the detected facial area. Using a geometric model of the detected eye area and facial area, the lip area is detected [12].

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, tongue and teeth, 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 1, and then the gray level [1], optical flow [13], and Sobel operator gradient of the pixel value of each grid are generated as features.

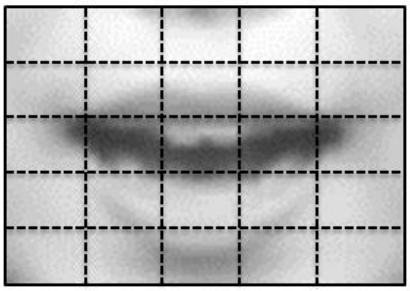

Fig. 1. Lip area overlaid with grid

The coordinate-based feature vectors are obtained by manually detecting the coordinates of 16 points that approximate the lip contours based on certain rules. The locations of the 16 points are shown in Figure 2.

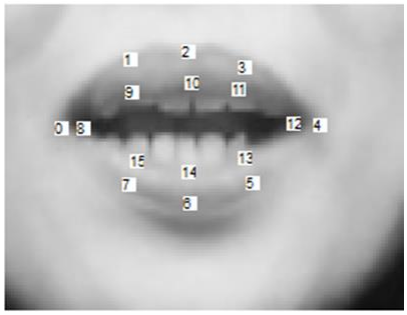

Figure 2. Coordinate-based features
(Outer lip width:0↔4, Inner lip width: 8↔12,
Outer lip height: 2↔6, Inner lip height: 10↔14)

Using the lip-shape coordinates detected by the above method, the feature vectors are comprised of the width/height ratio of the outer lip (ρ_1), the width/height ratio of the inner lip (ρ_2), the width of the outer lip (ρ_3), the height of the outer lip (ρ_4), the width of the inner lip (ρ_5), and the height of the inner lip (ρ_6).

4 Experiment and Results

A total of five subjects (three men and two women) were recruited for the study. The isolated words used in the experiment consisted of 하이갤럭시 /haigælʌksi/ (w1), 하이스마트폰 /haismɑ:rtpoʊn/ (w2), 하이카메라 (haikamera) (w3), 하이메시지 /haimesɪdʒ/ (w4), 하이카카오톡 /haikakaotok/ (w5), and 하이전화걸기 /haiʒʌnhwagʌlgi/ (w6). The five participants uttered each word three times, resulting in a total of 90 pieces of collected data. Dynamic Time Warping (DTW) was used to calculate the discrimination of various features extracted from the previously discussed lip image data.

The results of the recognition experiment for each feature generation method are shown in Figure 5. 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 w3 than the coordinate-based method.

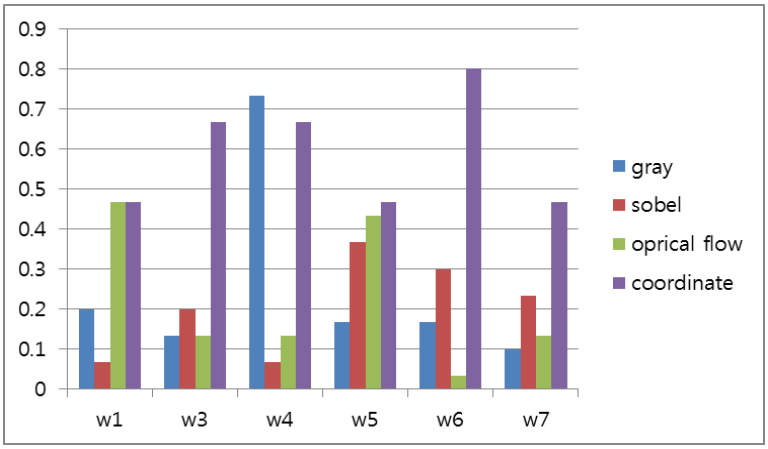

Fig 3. Mouth-shape recognition results by feature

Table 1. Mouth-Shape Recognition Confusion Matrix

	W1	W2	W3	W4	W5	W6	Recognition Rate
W1	7	3	1	1	2	1	46.7%
W2	2	11	1	0	1	0	73.3%
W3	0	1	10	3	0	1	66.7%
W4	2	2	3	7	0	1	46.7%
W5	1	1	1	0	12	0	80%
W6	2	3	0	2	1	7	46.7%

Table 1 is a confusion matrix of the optimal feature combination of the coordinate-based features, $\{\rho_1, \rho_6\}$. The recognition results showed high recognition rates for class w2 and class w5. These two classes showed higher recognition rates because class w2 was characterized as involving more closed-lip sounds, including /ㅁ/ and /ㅍ/, analogous to /m/ and /p/, respectively. Additionally, class w5 included the vowels /ㅗ/ (analogous to /o/), which are not included in other classes. In conclusion, the key features for determining utterances are the width/height ratio of the lips; however, the position of the articulators and change in overall movement of the lips are important for certain utterances.

5 Conclusions

The comparison of isolated word recognition rates through lip reading showed that higher average recognition rates were obtained from 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 showed higher recognition rates than the coordinate-based feature for certain isolated words, and the feature point using optical flow showed similar a recognition rate to the coordinate-based feature point. Thus, this requires additional analysis.

References

1. Chien, S.I, Bae, K.S.: “A study on the performance improvement of lipreading and speech recognition by fusing audio/visual information.” Technical Report. (2002) (in Korean)
2. Min, G.S.: “A Study on the Robust Lip Detection and LipReading System for Audio-Visual Speech Recognition in Mobile Environment.” Graduate School of Chonnam National University doctoral dissertation. (in Korean)
3. Lucey, S., Sridharan, S., Chandran, V.: “Adaptive mouth segmentation using chromatic features.” Pattern Recognition Letters, Vol. 23, No.11, pp.1293-1302 (2002)

4. Wang, S.L., Lau, W.H., Leung, S.H., Yan, H.: "A real-time automatic lipreading system." Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on. Vol. 2, IEEE (2004)
5. Lay, Y.L., Tsai, C.H., Yang, H.J., Lin, C.S., 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, pp.1465-1473 (2008)
6. Bagai, A., Gandhi, H., Goyal, R., Kohli, M., Prasad, T.V.: "Lip-reading using neural networks." International Journal of Computer Science and Network Vol. 9, No.4, pp.108-111 (2009)
7. Shaikh, A. A., Kumar, D. K., Yau, W. C., Che Azemin, M. Z., Gubbi, J.: "Lip reading using optical flow and support vector machines." In Image and Signal Processing (CISP), 2010 3rd International Congress on, Vol. 1, IEEE, pp. 327-330 (2010)
8. Skodras, E., Fakotakis, N: "An unconstrained method for lip detection in color images." In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, IEEE, pp. 1013-1016 (2011)
9. Ibrahim, M. Z., Mulvaney, D. J.: "Robust geometrical based lip-reading using Hidden Markov models." In EUROCON, 2013 IEEE, pp. 2011-2016 (2013)
10. Petajan, E.: "Automatic lipreading to enhance speech recognition." Illinois at Urbana-Champaign University Doctoral dissertation (1984)
11. Kim, Y.K., Lim, J.G., Kim, M.H.: "Lip Reading Algorithm Using Bool Matrix and SVM." International Conference on Small&Medium Business, pp267-268 (2015) (in Korean)
12. Chien, S.I., Choi, I.: "Face and facial landmarks location based on log-polar mapping." Biologically Motivated Computer Vision. Springer Berlin Heidelberg, pp.379-386 (2000)
13. Sun, D., Roth, S., Black, M.J.: "Secrets of optical flow estimation and their principles." Computer Vision and Pattern Recognition (CVPR) 2010 IEEE Conference on, pp. 2432-2439 (2010)

