Utility Analysis of Lip Features in Distinguishing Chinese Vowels and Lip Reading

Jianrong Wang¹, Yichao Zhang*¹, Wei Liu¹ and Yu Chen²

¹Department of Intelligence and Computing, Tianjin University, Tianjin, 300354, China
²School of Chinese Language and Culture, Tianjin University of Technology, Tianjin, 300384, China
*Corresponding author’s e-mail: zhangyichao@tju.edu.cn

Abstract. The lip region provides the most direct visual information in the process of multisensory speech perception, which is applied to speech recognition and lip reading. In this paper, we extract eight lip features in articulating the basic vowels [a], [e], [i], [u], [ü] in standard Chinese, and analyze the efficiency in distinguishing the five vowels combined with articulatory phonetics. We use Dense Convolutional Network (DenseNet) to process two-dimensional lip images and fuse the lip features to identify the Chinese with consonants. The results show that the application of lip shape features in Chinese vowel recognition and Chinese consonant lip reading is consistent. Two-dimensional lip images can effectively improve the recognition rate by fusing lip features in lip reading.

1. Introduction
Speech perception is a multi-sensory process which involves not only the perception of auditory information but also the visual information including the lip movements of speakers. Lips, as the most explicit speech organ in the face, provide the most direct and effective visual information of speech activities. This advantage enables the visual speech information acquired from the lip region to assist face recognition [1-2] and speech-driven face synthesis [3-4]. Due to the inimitability of the lips, the geometric features of lips play an important role in addressing the issue of speech recognition [5-6]. In addition, lip region features [7] and lip texture biometrics [8] are used in speaker recognition [9].

Considering the importance of lips [10] both in speech production and perception, articulatory phonetics has also paid considerable attention to the aspect of lip features in different language systems. To retrospect, researchers extracted the physiological feature vectors of lips, including the normalized width, height, average vertical distance and angle [11] to identify Persian [12], Malayalam [13] and Japanese [14-15]. As for the study of lip features in Chinese, Wang [16] and Yao [17] used the features such as the width and height of lip contour, the projection of the distance between the protrusions of the lips, to identify Chinese vowels. Considering that the transition from consonant to vowel in Chinese is very fast, Gang [18] selected the derivative of features like lip width, outer lip, inner lip line-height, protruding and inferior palatal process as new features of the speech flow. More specifically, Wu [19] proposed the use of lip protrusion, lip opening and lip shape area to distinguish between Chinese-specific round lip vowel [ü] and non-round lip vowels. Pan [20] proved that the lip width is the only physiological feature that can distinguish round vowels from non-round vowels in Standard Chinese.
In recent years, neural networks have become popular and significantly improved the performance of lip reading using multiple methods. The researchers apply convolutional neural networks (CNN) [21], deep neural networks (DNN) [22] to extract lip features. Taking into account the relationship between contexts in continuous speech recognition, researchers use recurrent neural networks (RNN) [23] and long short-term memory network (LSTM) [24].

In general, previous studies on Chinese and other languages using different neural networks have expanded the understanding of lip features and focused on improving the recognition rate of lip reading. Nevertheless, these studies do not discuss in-depth on the efficiency of these features in differing the vowels, especially how related vowels are distinguished from each other by these lip features. From distinguishing vowels to Chinese lip reading after adding consonants, whether the utility of lip features can remain the same, and whether there is any influence after adding consonants, these problems need to be studied.

Take above issues under consideration, this study applies a set of lip features to study the connection of the five basic vowels [a], [e], [i], [u], [ü] in Standard Chinese, and analyzes the efficiency of these features in distinguishing the five vowels. After that, we use DenseNet [25] to process two-dimensional lip image and then fuse the lip feature to identify the Chinese with consonant, and to analyze the utility of lip features in lip reading combined with articulatory phonetics.

The remaining sections are organized as follows: In section 2, we introduce the data description and data preprocessing. In section 3, we show the experiments and results. Finally, we discuss our research findings in section 4.

2. Data description and preprocessing

2.1. Subjects
Eight speakers, four males and four females, volunteer to participate in the experiment. All participants speak Standard Chinese as their first language and have no history of speech and hearing impairment.

2.2. Corpus
Bopomofo and four different pitch changes (tone 1-4) together determine the pronunciation of Chinese characters. In this corpus, we collect 6 vowels ([a], [o], [e], [i], [u], [ü]) and 7 consonants ([t], [f], [g], [p], [sh], [s], [x]), a total of 64 tokens. For example, four tuned syllables ([ā], [á], [ǎ], [à]) can be obtained by combining four tones (marked above the vowel) with atonal syllable [a], a total of 24 tokens. Various vowels and consonants are combined into polysyllables ([tā], [pò], [te], [tǐ], [tú], [fó], [guo], [xü], [sì], [shi]) with four tones (e.g. [tā], [tá], [tǎ], [tà]), a total of 40 tokens.

2.3. Data Collect
The experiment is conducted in a quiet room under normal light conditions. The experiment uses a fixed green curtain to standardize the image data. The hardware includes Kinect sensor with a sampling rate of 30 frames per second, a complete set of computer equipment. Speaker's lips keep level with camera (the fixed distance to Kinect camera is about 60 cm). The data recording environment is shown in ‘figure 1’. During the recording process, to expand the amount of data, each token is pronounced 10 times per person. Overall, this experiment collect 5120 (8 people × 64 tokens × 10 times) recording tokens. On average, the experiment lasts about an hour for one participant. The data recording process collect speech data, 2D color image data, depth image data, face 3D coordinate data.
2.4. Data Preprocessing

2.4.1. 2D-Image feature preprocessing. For 2D-image data, we use the open source OpenCV library to intercept a 128 × 100 lip region of interest. Considering the continuity of syllables pronunciation, we extract 16 consecutive frames (center − 8, center + 8) in the middle of the pronunciation to form a continuous sequence of image lip motion changes (4×4). This process is shown in ‘figure 2’.

2.4.2. Depth-Data feature preprocessing. For depth data, we use 3D coordinate data for 1347 face feature points to generate a 3D face model. Since the speaker’s head unconsciously shift during the recording (‘figure 3’ (a) (b) (c)), we rotate the face model with X-axis, Y-axis and Z-axes according to the symmetry of the face and the characteristics of the face. Finally, we get a standard three-dimensional face model (‘figure 3’ (d)) by calibration. We re-label the 1347 feature points of the face and select 160 feature points representing the lip area as shown in ‘figure 3’. Finally, we choose the 3D coordinate information of the 160 lip feature points in experiments.

After face calibration, we extract eight features including four ‘angles’ (upper angle, lower angle, left angle and right angle), ‘height’, ‘width’, ‘area’, ‘lip-pro’ from the 3D images from the 160 lip-points. As showed in ‘figure 3’, the feature ‘upper angle’ (the upper lip angle) is the cosine value of angle between the lip points 11, 1, 31. The feature ‘left angle’ (the left lip angle) is the cosine value of angle between the lip points 1, 11, 21. The feature ‘lower angle’ (the lower lip angle) is the cosine value of angle between the lip points 11, 21, 31. The feature ‘right angle’ (the right lip angle) is the cosine value of angle between the lip points 1, 31, 21. ‘Width’ is the distance between points 11 and 31. ‘Height’ is the distance between points 1 and 21. ‘Area’ is the area of the outer lip. ‘Lip-pro’ is the lip point change in front-back direction that is defined by the distance between the Kinect camera and the nearest lip point. It is noted that all the features are based on three-dimensional data, which can
effectively represent the depth information of lips and avoid the vulnerability of two-dimensional images to light, skin color, and normal head movements during the speaker's speech.

Considering the frame numbers are variable among tokens and vowels, this study regard the frame with the biggest height value as the representative frame of a given token. For each token, we put eight participants’ 80 feature data (8 people×10 times) together and impose an outlier analysis to remove the potential outliers. After wiping out the outliers, we calculate the average values of each participant’s feature data.

3. Experiment and result

3.1. Exp. 1

Vowels are the basis of pronunciation. In general, each Chinese Bopomofo contains 1 to 3 phonemes and must contain vowel phonemes. In view of the importance of vowels in Bopomofo, we analyze vowels in Exp. 1. The purpose of Exp. 1 is to explore the utility of lip-biometric features in distinguishing Chinese vowels, and analyze the result combing with Chinese phonetics knowledge.

Due to the limited variety of Chinese vowels, the distinguishing rate in the neural network is as high as 99% (we have confirmed), which is too close to compare the utility of the lip-biometric features. Therefore, two simple clustering algorithms, PAM and K-means (which have significant difference in different combination of features) are selected in Exp.1 to cluster the Chinese vowels using the lip-biometric features. The eight features have $2^8-1 = 255$ (2⁸-1) combinations after removing the empty set. We cluster analysis of 255 feature combinations.

The results of clustering the five vowels are shown in table 1 (only a few of the most recognizable combinations are shown). The results indicate that the discrimination rate was only 62.50% when all features were used, and it increase when some features were dropped in the clustering. When applying ‘width’ combined with ‘area’ or ‘left angle’, the clustering reach the highest discrimination rate (81.88%).

| Features                  | Discrimination rate |
|---------------------------|---------------------|
| All eight features        | 62.50%              |
| area                      | 48.75%              |
| width                     | 60.63%              |
| area, width               | 81.88%              |
| left angle, width         | 68.75%              |
| height, lip-pro, width    | 69.38%              |

Figure 4. The clustering plot of five vowels on PCA (features ‘area’ and ‘width’).
The result of clustering using ‘area’ and ‘width’ is shown in ‘figure 4’. In ‘figure 4’, each cluster is represented by the ellipse with the smallest area containing all its points. Each ellipse is labeled with the cluster name. Points in different ellipses/clusters have different shapes. Furthermore, we observed that the distribution of five vowels is not chaotic but regular. Thus, we decided to apply further clustering for those adjacent vowel pairs. For the vowel pairs that are well separated like [a] and [u], no further analysis is involved in this study. The results of clustering analysis for adjacent vowel pairs are shown in table 2. Only one combination of maximum discrimination rate is shown. We can see from the results that each combination of maximum discrimination rate contains ‘width’ feature.

| Method | PAM | Discrimination rate | Features | Discrimination rate | Features |
|--------|-----|---------------------|----------|---------------------|----------|
| vowels |      |                     |          |                     |          |
| [a] and [e] | right angle, width | 98.44% | height, lip-pro, width | 98.44% |
| [e] and [i] | upper angle, width | 82.81% | right angle, width | 82.81% |
| [i] and [ü] | width | 100% | width | 100% |
| [u] and [ü] | width | 95.31% | width | 95.31% |

We calculate the average recognition rate of each lip feature in distinguishing five Chinese vowels and adjacent vowel pairs, as shown in ‘figure 5’. From the average recognition rate, we can see that the ‘width’ is the best in distinguishing Chinese vowels.

![Figure 5. The average recognition rate of eight lip features based on two clustering methods.](image)

3.2. Exp. 2
In Exp. 2, we verify the utility of lip features on all data sets containing vowels and consonants. We use the DenseNet to evaluate lip features. The overall network is shown in ‘figure 6’.

![Figure 6. DenseNet Structure.](image)
Based on the vowel analysis results, we design several groups of comparative experiments in Exp. 2. Using the pre-processed two-dimensional lip splicing image as input as the benchmark. The remaining groups of experiments incorporate different lip depth features performed better in Exp. 1 (width, upper angle) and splicing images as input. Considering that the distinguishing effects of area, height, lip-pro are not obvious, a new ‘shape’ feature is generated by the ratio of height to width. The classification results are presented in table 3. Only the first three depth feature sets with the highest classification accuracy are shown.

| Feature Set        | Accuracy (%) |
|--------------------|--------------|
| Image              | 94.37        |
| Image +F\text{width} | 96.49        |
| Image +F\text{upper angle} | 95.68       |
| Image +F\text{shape} | 97.25        |

The results show that the classification accuracy is 94.37% with only the 2D lip splicing images. After 2D-image fusing with the depth feature, the accuracy of width feature set is higher than upper angle. The new shape fusion feature set achieve the best accuracy 97.25%.

4. Discussion

From the average recognition rate, we can see that the ‘width’ is the best in distinguishing Chinese vowels. In a previous study on the rounded vowels in Standard Chinese, Pan also found that the feature ‘width’ but not the feature ‘height’ can differentiate the rounded vowels from the non-rounded vowels effectively [20]. Similar to the results of that study, we also find that ‘width’ has the best capability of distinguishing the vowels in Standard Chinese while the feature ‘height’ contributes little, if any, to discriminate these vowels. This phenomenon reveals that ‘width’ and ‘height’ might not be coupled but independent with each other. In this study, we found that the two features have different change patterns: while the change of ‘height’ is monotonous (one rough peak in each wave), the change of the ‘width’ is much more complex (one or more than one rough peak/trough in each wave). This difference, we speculate, might cause the discrepancy between the two features in discriminating the vowels.

More interestingly, the four lip angles are different in discriminating the five vowels. Among the four lip angle features, ‘upper angle’ and ‘right angle’ performed well. By contrast, ‘left angle’ and ‘lower angle’ are not good features to differentiate the vowels. For the differences between ‘upper angle’ and ‘lower angle’, our result is in line with that of a study on Japanese vowel [14]. Since the facial muscles of controlling the lower lip are scarcer than that of the upper lip [19], the lower lip might be less mobile than the upper lip. As a result, the feature ‘lower angle’ is not as sensitive as the ‘upper angle’ in reflecting the differences among the five vowels. As for ‘left angle’ and ‘right angle’, due to unknown reasons, this study also reports that the two features have different efficiency in discriminating the five vowels.

Among the eight features, ‘area’ and ‘lip-pro’ are not good features in distinguishing the vowels. As a feature of reflecting the open-close degree of mouth, ‘area’ could not differentiate the five vowels well by itself, yet its performance is getting better when it cooperates with other features [19]. Meanwhile, in line with the previous study [19], this study finds that ‘lip-pro’ is not a good feature as well. Considering the feature ‘lip-pro’ is vulnerable to head movement which is unavoidable during speech production, its poor performance in vowel discrimination seems not that unreasonable.

For Chinese lip reading with consonants, after 2D-image fusion of depth features, the recognition rate of all feature sets has improved compared with the only image. The fusion width feature set still achieves good accuracy in more data containing vowels and consonants, consistent with the results in Exp 1. The ‘shape’ achieves the best recognition rate of 97.25%. The reason is that the ratio of height
to width reflects the area information, which makes ‘shape’ describes more specific lip changes during pronunciation.

The subsequent experiments not only apply fuse lip features to study the recognition of continuous speech sentences, but also extend the range of feature points to face. By studying the changes of organs and muscles in the range of face, it can be applied to face recognition and micro-expression recognition.

References

[1] Kumar, P.M., Gandhi, U., Varatharajan, R. (2019) Intelligent face recognition and navigation system using neural learning for smart security in internet of things. Cluster Computing 22: 7733–7744.

[2] Dewi, R., Sthevanie, F., Arifianto, A. (2019) Face expression recognition using Local Gabor Binary Pattern Three Orthogonal Planes (LGBP-TOP) and Support Vector Machine (SVM) method. Journal of Physics: Conference Series. 1192. 012048. 10.1088/1742-6596/1192/1/012048.

[3] Sadoughi, N., Busso, C. (2018) Expressive Speech-Driven Lip Movements with Multitask Learning. In: 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG). IEEE.

[4] Zhang, J., Fisher, R.B. (2019) 3d visual passcode: speech-driven 3d facial dynamics for behaviometrics. Signal processing, 160(JUL.): 164-177.

[5] Lin, B.S., Yao, Y.H., Liu, C.F., Lien, C.F., Lin, B.S. (2017) Development of novel lip-reading recognition algorithm. IEEE Access, 5: 794-801.

[6] Thein, T., San, K.M. (2018) Lip localization technique towards an automatic lip reading approach for Myanmar consonants recognition. In: International Conference on Information and Computer Technologies (ICICT), DeKalb, pp. 123-127.

[7] Nguyen, B. Tran, H. Dung, B. (2017) Robust lip feature detection in facial images. 867-871. 10.1109/FSKD.2017.8393389.

[8] Lu, Z., Wu, X., He, R. (2017) Person identification from lip texture analysis. IEEE International Conference on Digital Signal Processing. IEEE.

[9] Shang, D.P., Zhang, X.M., Xu, X.B., Peng, X.Y. (2018) Speaker Recognition Based on Lip-reading: An Overview. In: 7th International Conference on Network, Communication and Computing (ICNCC), Taipei, pp. 247-251.

[10] Maeda, S., Kurita, T., Kakita, Y. (1995) Physiology of the lips and modeling of lip gestures. Journal of Phonetics, 23(1): 243-254.

[11] Yargic, A., Dogan, M., (2013) A lip reading application on MS Kinect camera. In: IEEE International Symposium on Innovations in Intelligent Systems & Applications. Albena, pp. 1-5.

[12] Hosseini, M.M., Gharahbagh, A.A., Ghofrani, S. (2010) Vowel recognition by using the combination of haar wavelet and neural network. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Cardiff, pp. 331-339.

[13] Mohamed, F.K., Lajish, V.L. (2016) Nonlinear speech analysis and modeling for malayalam vowel recognition. Procedia Computer Science, 93, 676-682.

[14] Hisagi, M., Saitoh, T., Konishi, R. (2006) Analysis of efficient feature for Japanese vowel recognition. In: International Symposium on Intelligent Signal Processing and Communications, Tottori, pp. 33-36.

[15] Saitoh, T., Hisagi, M., Konishi, R. (2007) Analysis of features for efficient japanese vowel recognition. ICE Transactions on Information and Systems, 90-D(11), 1889-1891.

[16] Wang, M.J., Tian, J., Wang, X. (2010) Lip Contour Recognition Based on Context Information of Chinese Phoneme, Journal of Hebei University of Technology, vol. 39, no. 3, pp. 37-41.
[17] Yao, H.X., Lü, Y.J., Gao, W. (2002) Lip-Movement features extraction and recognition based on chroma analysis, Acta Electronica Sinica, vol. 30, no. 2, pp. 168-172.
[18] Gang, L.I., Wang, M.J., Ling, L. (2006) Improving Chinese lip-reading recognizing rate by unsymmetrical lip contour model, Optics & Precision Engineering, vol. 14, no. 3, pp. 473-477.
[19] Wu, Z. J., Lin, M.C. (1989) Summary of experimental phonetics. Higher Education Press, Beijing.
[20] Pan, X.S. Kong, J. P. (2014) A study of the round lip of Putonghua, Linguistics Series, vol. 2, 96-108.
[21] Vinyals, O., Toshev, A., Bengio, S., Erhan, D. (2015) Show and tell: A neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 3156-3164.
[22] Thangthai, K., Harvey, R., Cox, S., Theobald, B.J. (2015) Improving Lip-reading Performance for Robust Audiovisual Speech Recognition using DNNs. In: The 1st Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing, Vienna.
[23] Aulia, M., Maulana, R., Fanany, M. I. (2017) Sentence-level Indonesian Lip Reading with Spatiotemporal CNN and Gated RNN. 10.1109/ICACSIS.2017.8355061.
[24] Lu, Y.Y., Li, H.B. (2019) Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory. Applied Sciences. 9. 1599. 10.3390/app9081599.
[25] Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q. (2017) Densely Connected Convolutional Networks. In: CVPR. IEEE Computer Society, Honolulu, pp. 2261-2269.