Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 5476896, 9 pages
https://doi.org/10.1155/2020/5476896

Research Article

Chinese Tone Recognition Based on 3D Dynamic Muscle Information

JianRong Wang,[1,2] Li Wan,[1] Ju Zhang,[1] Qiang Fang,[3] Fan Yang,[1] and Jing Hu[1]

[1]*College of Intelligence and Computing, Tianjin University, Tianjin 300350, China*
[2]*Tianjin Key Laboratory of Advanced Networking, Tianjin University, Tianjin 300350, China*
[3]*Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China*

Correspondence should be addressed to Jing Hu; mavis_huhu@tju.edu.cn

Received 20 November 2019; Revised 13 February 2020; Accepted 11 April 2020; Published 31 May 2020

Guest Editor: Jianbiao Zhang

To advance the study of lip-reading recognition in accordance with Chinese pronunciation norms, we carefully investigated Mandarin tone recognition based on visual information, in contrast to that of the previous character-based Chinese lip reading technique. In this paper, we mainly studied the vowel tonal transformation in Chinese pronunciation and designed a lightweight skipping convolution network framework (SCNet). And, the experimental results showed that the SCNet was sensitive to the more detailed description of the pitch change than that of the traditional model and achieved a better tone recognition effect and outstanding antiinterference performance. In addition, we conducted a more detailed study on the assistance of the deep texture information in lip-reading recognition. We found that the deep texture information has a significant effect on tone recognition, and the possibility of multimodal lip reading in Chinese tone recognition was confirmed. Similarly, we verified the role of the SCNet syllable tone recognition and found that the vowel and syllable tone recognition accuracy of our model was as high as 97.3%, which also showed the robustness of our proposed method for Chinese tone recognition and it can be widely used for tone recognition.

1. Introduction

In recent years, the superior performance of lip reading in robust speech recognition has received widespread attention. The goal of lip reading is to improve the robustness of speech recognition in special situations such as low signal-noise ratio (SNR) or silent environments. However, due to the complexity and variability of Chinese pronunciation, the performance of lip-reading recognition in Chinese is not always satisfactory in real-world scenarios.

One of the most important tasks of lip-reading recognition is feature extraction. Currently, there are two main categories of visual information extraction in the lip reading system, i.e., pixel-based methods and model-based methods. Pixel-based methods extract visual features from the image directly or after some preprocessing and transformation. Yuhas et al. [1] used the greyscale image pixel information of the lip and its surrounding areas as features. Wolff et al. [2] used the horizontal and vertical scanning lines centred on the lips as the eigenvector. Since the method of directly using the pixel information of the image as a feature is blind, more effective and targeted approaches, such as discrete cosine transform (DCT), principle component analysis (PCA), singular value decomposition (SVD), discrete wavelet transform (DWT), and linear discriminant analysis (LDA) [3–5], were proposed to reduce the information redundancy. The pixel-based method can make full use of pixel information to extract more comprehensive lip features. However, the feature vectors are high dimensional and redundant. Also, the pixel-based method is very sensitive to light, shadow, pronunciation, and other conditions. Besides, model-based methods aim to establish a parametric mathematical model and then use the model parameters to describe lip contour

information. Kaynak et al. [6] used the horizontal and vertical distance of lip contours, the lip corner angle, and the first-order derivative of the lip corner angle. Zhang et al. [7] proposed geometric features of the lips, containing mouth width, upper/lower lip width, lip opening height/width, and the distance between the horizontal lip line and the upper. Model-based methods utilize low dimensional features to express image features, and the feature is typically not changed by factors such as translation, rotation, scaling, or illumination. Nevertheless, both methods extract relevant information directly from the region of interest (ROI) in the planar image [8].

With the development of high-sensitivity RGB-D cameras, the three-dimensional information of the speaker's face can be extracted more accurately. For instance, Yargıç and Muzaffer [9] developed a lip reading system that uses a Kinect camera to acquire the depth feature points and then extracts the angular features of the lip reading. Palecek et al. [10] studied the fusion performance of face depth data in isolated word visual speech recognition tasks. Rekik et al. [11, 12] proposed an adaptive lip-reading system based on image and depth data. Wang et al. [13] used 3D lip points obtained from Kinect, improving the performance of multimodal speech recognition. Studies by these pioneers have demonstrated the effectiveness of depth information in lip-reading recognition. Since the depth information is not affected by illumination, skin colour, etc. [14], the defects of the two-dimensional image information are compensated for. However, since the characteristics of the lips are usually obtained from discrete three-dimensional points or facial depth images, it is difficult to fully represent the characteristics of the lips.

The currently proposed lip-reading recognition based on 3D depth information does not consider the inherent texture problem of driving the lip motion during natural speech changes. In our previous work [15], to explore the internal mechanism of the speech process, we conducted an in-depth study on the facial texture information that drives the changes in lip reading and explored the facial texture information for lip movement changes in Chinese vowel pronunciation that have significant influence. However, since Chinese pronunciation is a strict tone-changing language, the transformation of the pitch has a significant role in the understanding of Chinese. Therefore, the exploration of Chinese tonal transformation in the current lip-reading research based on 3D information is important.

In this work, we focus on the study of the vowel tonal changes in Chinese pronunciation. Our main contributions are as follows. (1) For Chinese pronunciation tonal changes, we propose a new lightweight network framework, the SCNet, which is more sensitive to the transformation of details compared with the traditional network architecture. (2) We explore in detail the important influence of our proposed deep facial texture information on the change of vowel tones in auxiliary lip reading. (3) In syllable recognition with the depth texture, the experimental results show the ubiquity and good performance of the SCNet model in integrated tone recognition.

The rest of this paper is organized as follows. Section 2 introduces the data collection and preprocessing. Section 3 presents the proposed model architecture. Section 4 introduces our experimental results. Section 5 summarizes our work and introduces the future work.

2. Data Collection and Feature Preprocessing

2.1. Data Collection. Eight native speakers of Chinese, four males and four females, served as the subjects. All the subjects used standard Mandarin pronunciations without any accent influence. In the pronunciation of Chinese, each syllable has four different pitch changes (tones 1–4). In fact, there is a fifth pronunciation types in Chinese pronunciation, which is the unvoiced sound (i.e., a special silent tone in Chinese pronunciation) commonly spoken in Chinese. In order to explore the effects of different pitch transformations, we eliminated the unvoiced sounds that are rarely pronounced in Chinese, so in the experiment each syllable contained only one of the four commonly used tones. In terms of experimental data, we collected 5 vowels (/a/, /e/, /i/, /o/, and /u/) and 5 syllables (/ta/, /te/, /ti/, /fo/, and /tu/), a total of 40 tones. During the recording process, each tone was pronounced 10 times per person. For example, four tuned syllables (/ā/, / á/, /ǎ/, and /à/) were obtained by combining four lexical tones with the atonal syllable /a/.

The data acquisition device used a Microsoft Kinect V2 face real-time tracking camera and this camera through facial key points to generate real-time 3D point clouds (1347 facial key points). In [16], Mallick et al. have proved that the muscles of the facial expression recognition based on point cloud is successful, and it has been verified that the generation of 3D face point clouds is related to muscle distribution. At the same time, their experiments show that the shape of the face of point cloud generated face has nothing to do and can be very stable in different faces of the same position. Meanwhile, [17, 18] also prove the stability and effectiveness of Kinect V2. To ensure its quality, we collected the data in a standard silent room. The data collection scenario is shown in Figure 1.

During the process, we reindexed the 1347 points. The index of feature points in the lip area is shown in Figure 2(b), which used only the collected image information and 3D depth information. By considering the changes in the head model during movement, we corrected the head rotation angle in the $X-$ axis, $Y-$ axis, and $Z-$ axis directions. As an example, the angle between vector $\overrightarrow{P_{11}P_{31}}$ (P_{11} and P_{13} are two points in Figure 2(b)) and plane XY is calculated as follows:

$$\alpha_{XY} = -1 \times \arctan\left(\frac{(Z_{31} - Z_{11})}{(x_{31} - x_{11})}\right), \tag{1}$$

where (x_{11}, y_{11}, z_{11}) and (x_{31}, y_{31}, z_{31}) are the coordinates of P_{11} and P_{31} and 31 and 11 represent the coordinate point numbers on the plane XY. The rotated face point coordinates parallel to the XY plane are constructed by the following algorithm.

Figure 1: Kinect V2 recording data experimental scene.

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} \cos(\alpha_{XY}) & 0 & \sin(\alpha_{XY}) \\ 0 & 1 & 0 \\ -\sin(\alpha_{XY}) & 0 & \cos(\alpha_{XY}) \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \quad (2)$$

Finally, we acquired the standard point set of the real speaker's face.

2.2. Feature Preprocessing

2.2.1. Image Feature Preprocessing. For the collected image information, we used the open source OpenCV lib library to intercept a 128×100 lip region of interest, as shown in Figure 4(a), and then used the image sequence representation method proposed by Saitoh et al. The pronunciation of the syllables extracts 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, from left to right, top to bottom) and uses a gamma transform ($V_{\text{out}} = V_{\text{in}}^{\gamma}$) for light enhancement to augment the data, as shown in Figure 4(b) (take 16 sheets and then sort).

2.2.2. Muscle Dynamics Features. According to this study, there are six main types of muscles that drive lip movement in facial muscles. The distribution of the facial functions and characteristics of each muscle are presented in Tables 1 and 2 reflect the specific names of each muscle and the characteristic point identification of each muscle in the kinect data. In the specific depth texture feature representation, we extracted the two most representative depth, muscle length change, and muscle dynamic characteristic data points.

(1) Muscle Length Change Information. The length feature is expressed as $[1/R]$, where l represents the muscle length vector at the time of speech and R represents the muscle length vector at the time of relaxation, which eliminates the differences between different speakers.

(2) Muscle Dynamics Information. The muscle dynamics information characterizes the relationship between the facial muscles and facial feature points and reflects the intrinsic commonality between different speakers. We also analysed the effects of different muscles on the displacement of the feature points as the drivers of muscle dynamic transformation. Regarding the feature information, the vector variation between the muscles is obtained by calculating the transformation trend of different feature points in adjacent frames. The specific expression is as follows:

$$F_{\text{muscle\_}i} = \left[\frac{P_{j\text{–end}} - P_{j\text{–start}}}{l_j} \right] \cdot \overline{V}_{\text{muscle}}, \quad (3)$$

where $F_{\text{muscle\_}i}$ represents the momentum change of the feature point i, $P_{j\text{–start}}$ and $P_{j\text{–end}}$ represent the start and end points, respectively, of the muscle j, and the direction of the muscle movement at each point is represented by decomposing the displacement subvector of each point. $\overline{V}_{\text{muscle}}$ Indicates the length of movement of each muscle point.

3. Network Architecture

Considering the subtle differences in the mouth shape changes in Chinese tonal changes, we designed a lightweight skip convolutional structure network (SCNet) with subtle descriptions of feature changes to evaluate our proposed 3D lip features and to explore the feasibility of tonal changes and syllable lip-reading recognition. The overall architecture is shown in Figure 3.

The network architecture was inspired by that of VGG [19] and ResNet [20]. In the initial phase of the network, we used three 3×3 convolutional layers with a stride of 2 to extract the surface features of the image. This network structure reduces not only the overall parameters of the network but also the accuracy loss of the feature map.

The main body of network structure is two connected feature extraction blocks, and they different from the current remaining block structure. Two subconnection blocks adopt different subsampling expressions. At the back of block 1, to make the edge features more obvious, the maximum pool was used to indicate the specificity of different features, highlighting the features of different feature maps. And, at the block 2, to make the features, the map was associated with the specificity of the feature maps more smoothly and effectively using global average pooling. The two connection block structures in the frame were slightly different. In the second block, to maximize the smoothing effect after block 1, the last convolutional layer output channel in block 2 was doubled and the rest was the same as that of block 1. This structure also showed good performance in the experiment. At the last end, a 128-dimensional linear layer was connected, and then the classification probability was obtained.

3.1. Skip Convolution Structure.

We used a skip connection in each block. The structure of each block is shown in Figure 2(b), and the connection of each block is defined as follows:

$$y = F(x) + G(x), \quad (4)$$

where x and y represent the input and output, respectively, of each block and $F(x)$ represents the learning function of

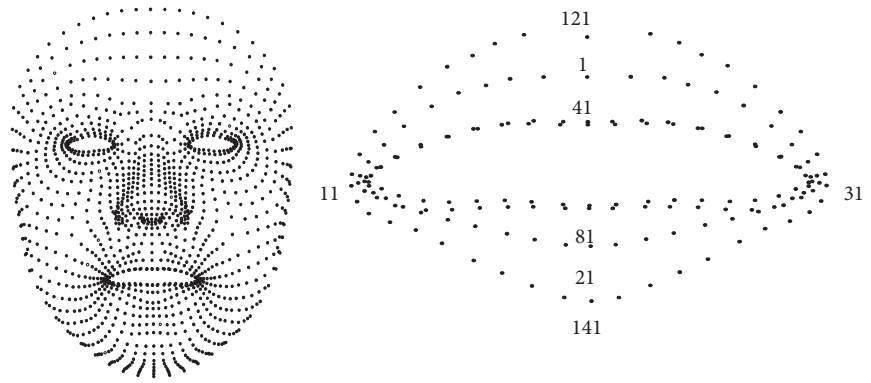

Figure 2: (a) Predefined 1347 planar facial points. (b) Reindexed 160 points of lip area.

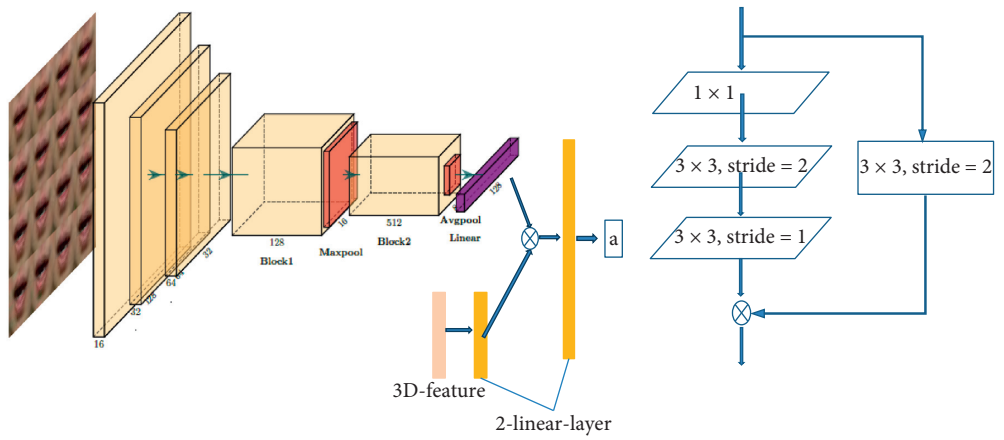

Figure 3: Our SCNet structure. (a) The overall structure of the model and (b) the skip connection structure.

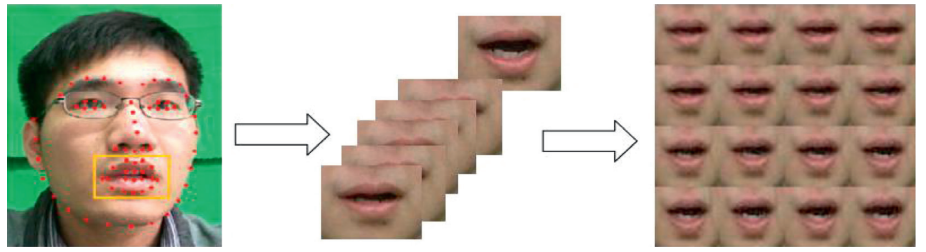

Figure 4: Picture-stitching process. (a) Feature extraction of interest and (b) representation of the image sequence splicing process.

Table 1: The structures and functions of the major muscles.

Muscle name	Structure	Function
Levator labli superioris	From the medial infraorbital margin to the skin and muscle of the upper lip	Elevates the upper lip
Levator anguli oris	From the canine fossa, below the infraorbital foramen	Draws the angle of the mouth
Zygomaticus	Extends from the zygomatic arch to the corners of the mouth	Draws the angle of the mouth
Buccinator	From the alveolar processes of the maxilla and mandible and the temporomandibular joint	Pulls back the angle of the mouth
Orbicularis oris	Composed of four independent quadrants, gives an appearance of circularity	Encircles the mouth
Depressor anguli oris	From the tubercle of the mandible to the modiolus of the mouth	Depresses the angle of mouth

Table 2: The starting and ending coordinates of each muscle.

Muscle name	Start point	End point	Affected lip points
Levator anguli oris	603	126	125, 126, 127, 128
Zygomaticus	650	131	129, 130, 131
Buccinator	522	131	127, 128, 129, 130, 131
Levator labli superioris	769	165	125, 126
Orbicularis oris	717	126	125, 126
Depressor anguli oris	665	127	127, 128, 129, 1230, 131

direct connection. As Figure 3(b) shows, the direct connection is composed of three convolution layers, so $F(x)$ is specifically expressed as $F(x) = W_3 \cdot \sigma(W_2 \cdot \sigma(W_1 \cdot x))$, in which σ is LeakyReLU and $G(x)$ is the skip connection, which represents the connection structure of a layer and is given by the formula $G(x) = W \cdot x$. Since the regularization layer was introduced, to reduce the parameter changes in this architecture, the bias item was not led into. Finally, the $F(x) + G(x)$ operation represents the direct weight addition of the direct and skip connection, rather than the corresponding result splicing.

Equation (4) is mainly divided into two parts: direct connection structure and skip structure. In the stage of direct connection structure, first we used a 1×1 convolution, followed by a 3×3 convolution, with a stride of 2 to obtain more detailed feature information, and then the network optimization is connected to a 3×3 convolution kernel, with a stride of 1 to simulate the processing of the Sobel matrix on the feature boundary. This structure makes the boundary features more obvious, so that the feature was better characterized in the feature judgement area. In the skip module, we used a 3×3 convolution block, with a stride of 2, and the number of channels was increased. This procedure generates the same channel for the network, and the same size is more convenient for feature stitching. This method also ensures the fusion of the image on the feature structure. The purpose of the traditional Res block is to ensure the characterization of the local structure and the global feature to make the network structure more representative. We use this structure to consider that the 1×1 convolution has retained the global feature, using a 3×3 convolution. This convolution ensures the multiscale representation of the network structure.

3.2. Feature Fusion Structure. The expression for feature fusion structure is given as follows:

$$\text{Infor}_{\text{cat}} = F_{\text{fusion}}\left(\text{Infor}_{\text{img}}, \text{Infor}_{\text{depth}}\right). \tag{5}$$

To better integrate the depth information and picture information, we adopted a decision fusion method to deeply integrate the two different kinds of information. The specific expression is shown in formula (5), where $\text{Infor}_{\text{img}}$ represents the 128-dimensional information acquired by the SCNet. The depth feature, $\text{Infor}_{\text{depth}}$ represents the depth feature of the shallow stitching after two layers are fully connected, and F_{fusion} indicates the fusion strategy. Thus, the feature, $\text{Infor}_{\text{cat}}$, after the fusion of the two, was decoded by a linear layer of one layer and output.

3.3. Implementation Detail. In the experiment, the input size of our image is 112×112. Since the image was adjusted before input, no corresponding data enhancement method was used during the experiment. Batch normalization (BN) [21] was adopted in the network after each convolution, before activation and after the BN. For the network weights, the random initialization method was adopted and the network was trained from zero. An Adam optimizer was used in the experiment, and the small batch size was set to 30. The learning rate started at 0.0003, and the expression of the learning rate attenuation functions is shown in the following formula:

$$\text{new\_lr} = \text{lr} \times \gamma^{(\text{epoch} - \text{sleep}_{\text{epoch}}+1)/\text{half}}, \tag{6}$$

where lr represents the last round of the learning rate, $\text{sleep}_{\text{epoch}}$ (20) iterations decay once, and each damping coefficient is γ (0.5) times (epoch − sleep_epoch + 1)/half (5) − th. We did not use dropout during the implementation.

4. Experiments and Results

In the experiment, to verify the smoothness of the proposed model on the whole dataset, we set the experimental scheme to a five-fold cross-validation and calculated the average of all the results as the final experimental result.

4.1. Cross-Validation. To ensure the full use of the data and the accuracy of the experimental results in our experiments, we designed a 5-fold cross-validation. We randomly divided all the experimental data into 5 parts. Water sampling was used for the data division. The data in each sample set consisted of only 1860 groups. Four tests were used to train one test, and the experiment was performed for a total of 5 rounds, so that each could be used as the training set and test set and each experiment would give an independent result.

Because vowels play a leading role in the whole pronunciation process, in the experiment, in order to verify the difference between the entire syllable recognition effect and the different syllable recognition performance of each syllable, we first aimed at each vowel recognition accuracy was discussed, and then further analysis of tone recognition of vowels with different tones. By using different speech expressions, we ignore the unvoiced sounds in Chinese pronunciation to verify that our proposed SCNet has considerable experimental results in terms of accuracy of tone recognition and accuracy of the entire syllable recognition.

4.2. Vowel Detection and Vowel Tone Detection. We first verified the validity of our proposed model and compared it with the traditional models (VGG, ResNet, DenseNet [22]); in addition, we tested the effects of the different models on vowel recognition and vowel tone recognition. To ensure the fairness of the comparison, a linear 1000×128 layer and a softmax classification layer were added to the traditional model, and the optimal values the parameter settings were selected.

Figures 5 and 6 show the single vowel recognition results and the vowel tone recognition results, respectively. By comparing the two images quantitatively, we found that all the models showed good recognition performance; specifically, the proposed vowel distinction SCNet reached a recognition rate of almost 100%, and the tone recognition effect was significantly higher than that of the traditional model structure. A comparison of the overall results of several models in terms of the network depth, parameters, and accuracy is shown in Table 3. It was found that the SCNet gave the optimal values of the three parameters, especially those of the parametric variables. Compared with those of the previous models, the SCNet parameters were only 1/50 of the VGG value, 1/4 of ResNet value, and 1/3 of DenseNet value and even more advantages of the experimental results. These results indicated that our designed model was advantageous for processing real-time data and had better performance than that of the existing traditional framework.

Our analysis of this experimental phenomenon is based on the application of the SCNet architecture to the transformation of subtle differences in the datasets. This architecture showed good results for the description of the data details.

As a whole, the experiment can show such excellent results and attribute the success to the following characteristics of the network structure: (1) in tone recognition, the degree of differentiation of the mouth shape between different tones of the same syllable is very small, and we used a $3\times$ filter in the experiment. The use of such a small convolution kernel can enhance the fine feature structure discrimination. (2) Based on several previous verifications, it was proved that skipping convolutions can preserve the feature transformations between feature maps, which in addition is more conducive to the propagation of gradients than are traditional direct connections. The jump connection proposed in this paper showed that our method can capture more delicate network structure features and thus improve the fine discrimination performance. (3) Different downsampling methods between different structural blocks can be used in feature selection, highlight the propagation between different features, and make the network structure smoother, which is more conducive to the expression of different detailed features.

4.3. Texture Depth Information Fusion. To better verify the validity of the depth texture information in tone recognition, we designed a series of experiments to confirm the correctness of our conjecture.

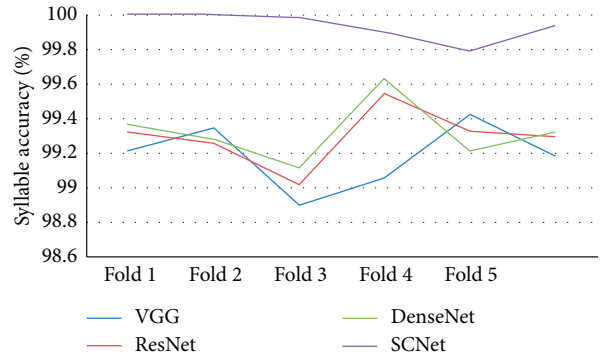

Figure 5: Vowel recognition results of different models.

The results of the tone recognition of the picture only and the tone recognition after the fusion of the depth information are shown in Figure 7. The experimental results showed that after fusion of the texture depth information, the recognition result of the image-only tone recognition increased by 2%, and especially in the case of low picture recognition rate, the effect on the tone recognition was obvious, which indicated that our proposed 3D depth texture information significantly influenced the auxiliary tone recognition. This effect occurred because image-based features are not sufficient to fully represent continuous lip motion. The feature tone recognition of colour images is sensitive to light, speaker skin colour, and camera acquisition quality. However, 3D information has good anti-interference for this kind of disadvantage and is hardly affected. Our proposed facial texture depth information largely compensates for the defect of lip pronunciation in tone recognition caused by environmental problems and complements the image-only lip pronunciation method.

Figure 8 shows the results of the model recognition for adding different noise types. In the experiment, the random Gaussian noise with the mean $\in [0, 10]$ and variance $\in [10, 20]$ was added to simulate the recognition scenario for different photographic definitions, and the gamma algorithm with the gamma interval $\in [1, 8]$ was used to adapt to changes in the lighting due to real-life changes. Adding such dynamic noise can better reflect the robustness of different models in natural scenes and the ubiquitous ability of different frameworks. Unexpectedly, the performance of the proposed SCNet model was much higher than that of the traditional model, which shows that our framework has better application performance in real-world scenarios. Similarly, for the performance of the recognition effect before and after the texture depth information, there was a stable improvement effect of more than 0.5% after the fusion of the depth information, indicating that the fusion depth information is more meaningful for the recognition of the real scene.

4.4. Syllable Recognition. Since tone change occurs in all Chinese pronunciations and the consonant is attached to the vowel, the difficulty of syllable recognition is greater than that of the vowels. To further verify the effectiveness of our

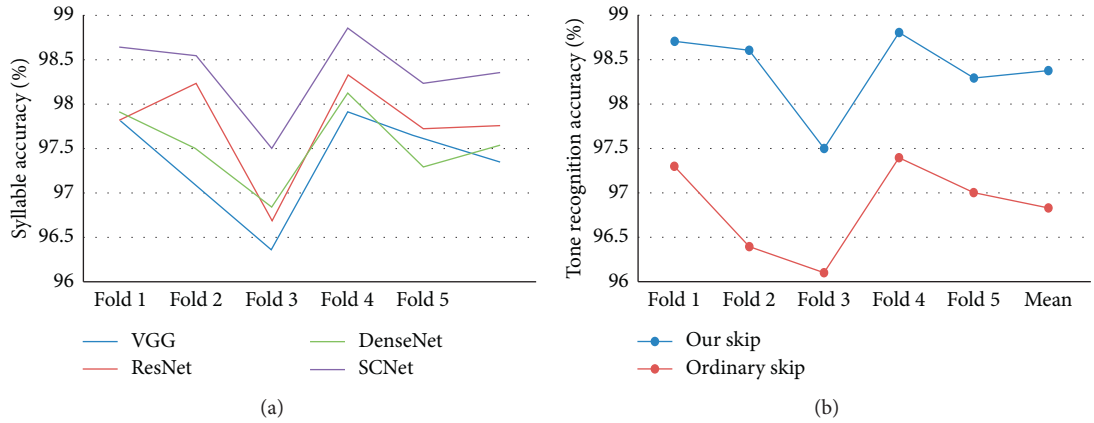

Figure 6: Vowel tone results of different models. The influence of (a) different methods and (b) different skip connections on the accuracy of vowel recognition.

Table 3: Comparison of the network depth, parameters and experimental accuracy of the four different models.

Method	Depth	Params	Accuracy
VGG	11	531.5M	97.35
ResNet	18	46.9M	97.75
DenseNet	121	33.4M	97.528
SCNet	10	10.9M	98.352

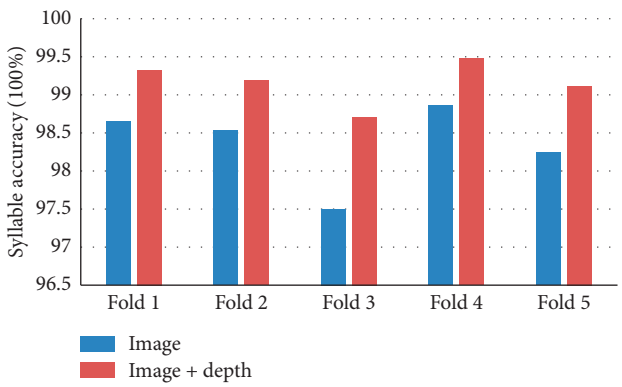

Figure 7: Convergence depth information and comparison of image-only results.

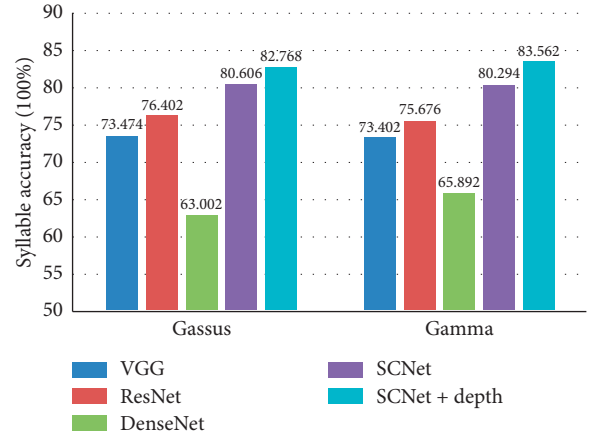

Figure 8: Robustness test of several different dynamic noise models.

proposed SCNet in the recognition of all Chinese tones, we also verified the performance of the model in the recognition of 40 mixed tones based on 5 vowels (/a/, /e/, /i/, /o/, and /u/) and 5 syllables (/ta/, /te/, /ti/, /fo/, and /tu/).

The recognition results are shown in Figure 9. Although the pitch recognition of syllables is more difficult according to the theory, our SCNet model was robust, and a high recognition rate of 97.364% was obtained, indicating that our model had not only a good vowel tone recognition performance but also an excellent Chinese tone recognition performance. Moreover, after adding the depth texture information, the average recognition result of the pitch showed a 0.2% improvement. Since the pronunciation of the syllable is more complicated than that of the vowel and the pronunciation organ is more involved, the facial depth may be relevant. Texture information has a greater impact on the recognition of syllables. A comparison with our previous conjectures indicates that deep texture information has a very clear effect on the recognition of the Chinese lip to assist in lip reading for both consonant and vowel tone recognition.

5. Summary

This work was mainly focused on the difficulty of tone recognition in Chinese lip-reading recognition. In this paper, we designed an efficient lightweight network framework, SCNet, based on a comprehensive and effective lip-

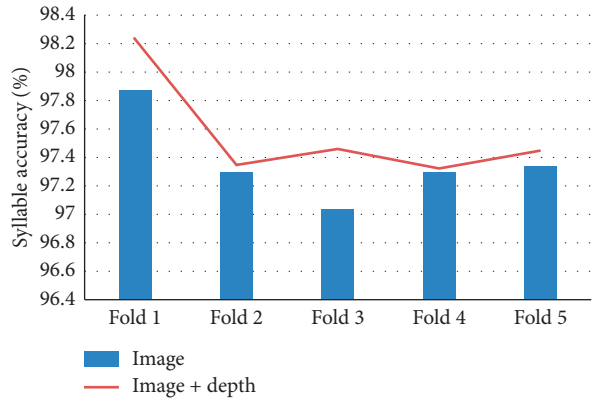

Figure 9: Tone recognition results of vowels and syllables.

reading feature extraction method and verified the effectiveness of our proposed network framework by several experiments. In the study, we carried out an in-depth verification on the proposed framework. Comparison experiments showed that the framework can accurately identify the tones of Chinese pronunciation. In addition, the facial texture depth information and picture information fusion demonstrated the feasibility of facial texture depth information to help the recognition of Chinese tones.

With the wide application of depth cameras on video equipment, lip reading will better assist speech recognition in the future and improve the robustness of speech recognition in different environments. The dataset used in this paper consisted of independent syllables, but the results show that the proposed method is practical and can be effectively applied to future large-scale datasets.

Data Availability

The data used to support the findings of this study are available from the first author upon request.

Conflicts of Interest

The authors declare no potential conflicts of interest with respect to the authorship and/or publication of this article.

Acknowledgments

This study was financially supported by the National Natural Science Foundation of China (grant no. 61977049) and by the Tianjin Key Laboratory of Advanced Networking.

References

[1] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, vol. 27, no. 11, pp. 65–71, 1989.

[2] G. J. Wolff, K. V. Prasad, D. G. Stork, and M. E. Hennecke, "Lipreading by neural networks: visual preprocessing, learning, and sensory integration," in *Advances in Neural Information Processing Systems*, Morgan Kaufmann Publishers Inc. Burlington, MA, USA, 1993.

[3] P. Scanlon and R. Reilly, "Feature analysis for automatic speech reading," in *Proceedings of the IEEE Fourth Workshop on Multimedia Signal Processing*, Cannes, France, October 2001.

[4] P. S. Aleksic and A. K. Katsaggelos, "Comparison of low- and high-level visual features for audio-visual continuous automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics*, Montreal, Canada, May 2004.

[5] I. Matthews, G. Potamianos, C. Neti, J. Luettin, and A. Ascom Systec, "A comparison of model and transform-based visual features for audio-visual lvcsr," in *Proceedings of the IEEE International Conference on Multimedia & Expo*, Tokyo, Japan, August 2001.

[6] M. N. Kaynak, Q. Zhi, A. D. Cheok, K. Sengupta, Z. Jian, and K. C. Chung, "Analysis of lip geometric features for audio-visual speech recognition," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 34, no. 4, pp. 564–570, 2004.

[7] X. Zhang, R. M. Mersereau, and M. A. Clements, "Audio-visual speech recognition by speechreading," in *Proceedings of the International Conference on Digital Signal Processing*, Orlando, FL, USA, May 2002.

[8] J. Bin, Y. Jiachen, L. Zhihan, T. Kun, M. Qinggang, and M. Yan, "Internet cross-media retrieval based on deep learning," *Journal of Visual Communication & Image Representation*, vol. 48, pp. 356–366, 2017.

[9] A. Yargıç and D. Muzaffer, "A lip reading application on MS Kinect camera," in *Proceedings of the IEEE INISTA*, Albena, Bulgaria, June 2013.

[10] K. Palecek, *Extraction of Features for Lip-Reading Using Autoencoders*, Springer, Berlin, Germany, 2014.

[11] A. Rekik, A. Ben-Hamadou, and W. Mahdi, *A New Visual Speech Recognition Approach for RGB-D Cameras*, Springer, Berlin, Germany, 2014.

[12] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "An adaptive approach for lip-reading using image and depth data," *Multimedia Tools and Applications*, vol. 75, no. 14, pp. 8609–8636, 2015.

[13] J. Wang, J. Zhang, H. Kiyoshi, W. Jianguo, and D. Jianwu, "Audio-visual speech recognition integrating 3D lip information obtained from the Kinect," *Multimedia Systems*, vol. 22, no. 3, pp. 315–323, 2016.

[14] J. Yang, B. Jiang, B. Li, K. Tian, and Z. Lv, "A fast image retrieval method designed for network big data," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2350–2359, 2017.

[15] J. Wei, F. Yang, J. Zhang, R. Yu, M. Yu, and J. Wang, "Three-dimensional joint geometric-physiologic feature for lip-reading," in *Proceedings of the 2018 IEEE 30th International Conference on Tools with Artificial Intelligence*, Volos, Greece, November 2018.

[16] T. Mallick, P. Goyal, P. P. Das, and A. K. Majumdar, "Facial emotion recognition from kinect data—an appraisal of kinect face tracking library," in *Proceedings of the International Conference on Computer Vision Theory and Applications*, Rome, Italy, February 2016.

[17] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, 2012.

[18] The Difference between Kinect v2 and v1, 2020, https://skarredghost.com/2016/12/02/the-difference-between-kinect-v2-and-v1.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[20] S. R. K. He, X. Zhang, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 770–778, Las Vegas, NV, USA, July 2016.

[21] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on Machine Learning (ICML)*, Lille, France, July 2015.

[22] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, Honolulu, HI, USA, July 2017.