Gabor Based Lipreading with a New Audiovisual Mandarin Corpus

Yan Xu, Yuexuan Li, and Andrew Abel(✉)

Research Institute of Big Data Analytics (RIBDA),
Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
{Yan.Xu,Andrew.Abel}@xjtlu.edu.cn,
Yuexuan.Li15@student.xjtlu.edu.cn

Abstract. Human speech processing is a multimodal and cognitive activity, with visual information playing a role. Many lipreading systems use English speech data, however, Chinese is the most spoken language in the world and is of increasing interest, as well as the development of lightweight feature extraction to improve learning time. This paper presents an improved character-level Gabor-based lip reading system, using visual information for feature extraction and speech classification. We evaluate this system with a new Audiovisual Mandarin Chinese (AVMC) database composed of 4704 characters spoken by 10 volunteers. The Gabor-based lipreading system has been trained on this dataset, and utilizes the Dlib Region-of-Interest(ROI) method and Gabor filtering to extract lip features, which provides a fast and lightweight approach without any mouth modelling. A character-level Convolutional Neural Network (CNN) is used to recognize Pinyin, with 64.96% accuracy, and a Character Error Rate (CER) of 57.71%.

Keywords: Audiovisual · Speech recognition · Chinese · Gabor transform

1 Introduction

Human speech is multimodal, with audio and visual information used both in the perception and production of speech. This relationship has been heavily investigated in the literature, with a detailed summary in Abel and Hussain [2]. Lipreading is an approach that interprets lip movement [13], inspired by human cognitive abilities. This can be used in speech recognition, identity recognition, human-computer intelligent interfaces and multimedia systems. Many proposed lipreading systems have very high recognition rates, for instance, the LipNet [4] and 'Watch, Listen, Attend, and Spell' (WLAS) [12] systems. However, most of them use an English corpus such as Grid. Chinese is spoken by around one fifth of the world's population, but written Chinese does not indicate its pronunciation from its character. However, "Pinyin" can be used to mark the pronunciation of Chinese Mandarin [6]. Lipreading is a difficult task to apply to Pinyin, with difficult to identify ambiguities, like similar lip shapes (e.g. 'p' and 'b'), liquids and

J. Ren et al. (Eds.): BICS 2019, LNAI 11691, pp. 169–179, 2020.
https://doi.org/10.1007/978-3-030-39431-8_16

nasals (e.g. ‘n’ and ‘l’, are similar to ‘ni’ and ‘li’), blade-alveolars and retroflexes (e.g. ‘ci’ and ‘chi’), and front and back nasal sounds (e.g. ‘nin’ and ‘ning’).

In this paper proposes an improved Chinese lipreading system to recognize Pinyin and tone in Chinese speech. To evaluate this, we introduce a new AVMC database, composed of 4704 Pinyin words spoken by 10 volunteers in a clean visual and acoustic environment. For feature extraction, an improved fast and lightweight architecture based on Gabor transforms was developed. A CNN was used to test these lip features, with a character-level performance of 64.96%.

2 Related Work

Traditionally, lipreading has two key components: feature extraction and recognition. Recently, some end-to-end lipreading systems have been developed, such as LipNet [4], and Long Short-Term Memory (LSTM) systems [17]. Chung et al. [12] proposed a character-level architecture called WLAS, with a 3% Word Error Rate(WER), Weng [18] achieved 82.0% word recognition rate for the LRW corpus, and Petridis et al. [11] obtained a 94.7% recognition rate for OuluVS2. These end-to-end approaches use an image directly to self-learn features, and are difficult to explain. We are more interested in extracting lip features, as they can be used for more than just model results. Classic methods are arguably more lightweight and explainable [3]. These include Active Appearance Models (AAM), Discrete Cosine Transform (DCT), and Gabor Wavelet Transform (GWT). DCT is good at concentrating energy into lower order coefficients, but is sensitive to illumination changes, and is difficult to form an intuitive understanding of [1].

Gabor features are insensitive to variance in illumination, rotation, and scale. They can focus on facial features such as eyes, mouth and nose, with optimal localization properties in both spatial and frequency domains [5]. We propose GWT to extract lip features, which is a fast and lightweight approach without any mouth modelling. Compared to deep learning models, GWT reduces speech recognition training time, and is more suitable for small datasets. Sujatha and Santhanam [14] used GWT to correct mouth openness after using a height-width model when extracting 2D lip features, with word recognition of 66.83%. We extract seven lip features, which includes six 2D features and one 3D feature. Hursig et al. used Gabor features to detect the overall lip region [10], while we obtain detailed features. Dakin and Watt proposed that horizontal Gabor feature performance is good for facial feature recognition by using Gabor filters of different orientations [7], and was implemented by the authors [1]. Here, we present an improved visual feature extraction system.

3 A New Audiovisual Mandarin Speech Corpus

Many English-language audiovisual speech corpora have been published, including LRS [12], AVLetter, AV-TIMIT, CUAVE, Grid,OuluVS and XM2VTSDB [20]. However, Chinese lipreading research suffers from a shortage of published and available corpora. Zhang et al. in 2009 collected a large-scale Chinese corpus by recording CCTV news broadcasts, which includes 20495 natural Chinese sentences [19]. However, there is a complex visual background and a noisy speech environment. This paper introduces a new audiovisual Chinese corpus recorded in a clean environment. Chinese is a tonal language, consisting of individual characters, each of which have an initial, a final, and a tone associated with them. To perform accurate initial speech recognition and further analysis, we require a labelled video corpus of distinct Chinese characters, recorded in a clean environment, and we therefore created the AVMC dataset.

Data Acquisition: 162 Chinese characters were collected from the general specification table published by the Chinese ministry of education. These characters are chosen with a reasonable distribution of initials, finals and tones. 10 native Mandarin Chinese volunteers were used (see Fig. 1). The data acquisition procedure for each volunteer was: (1) sign participant information and consent form; (2) read caption list; (3) practice recording for 1–2 min; (4) record for all captions and repeat 3 times. During recording, volunteers were asked to pause between each word, and if they made mistakes, paused and repeated. Mistakes not identified during recording were identified later in the editing process. This produced 30 videos, each being a volunteer reciting all 162 characters in a quiet environment, with a plain blue screen as background. To ensure they were looking directly at the screen, a teleprompter was used. As some Chinese characters have the same pronunciation, there are in total 158 types of pronunciations including both correct and wrong utterances. The video was recorded at a resolution of 1920 × 1080, at 50 fps, and the audio was recorded at 48 kHz.

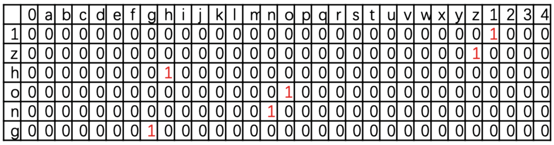
Fig. 1. Example frames of all speakers in AVMC

Data Pre-processing: The 30 raw videos were edited using Adobe Premiere Pro to remove silent pauses, breaks, coughs, off-camera noise, and maintain a consistent order of characters. The processed videos are stored in mp4 format.

Data Labeling: The pauses were manually identified, and video captioning software (Arctime) was used to label initial, final, tone and Pinyin. The occasional vocalized mistakes were not corrected but labelled with actual pronunciations. In a character level representation, the classes for initials, finals, and tones are given shown as follows:

- initials: 0, b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, w, x, y, z
- finals 0, a, e, g, i, n, o, r, u and v
- tones 1, 2, 3, 4

For lipreading, the model uses a distinct Pinyin character representation, as shown in Fig. 2. Here, the Pinyin for zhong with the first tone would be represented as having the characters ‘1’, ‘z’, ‘h’, ‘o’, ‘n’, ‘g’. However, Pinyin such as ‘ban’ with second tone would be shorter and could be represented as ‘2’, ‘b’, ‘a’, ‘n’, ‘0’, ‘0’. Here, 0 means there is no character in this location.



	0	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	1	2	3	4
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
h	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
g	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 2. Character level representation of Pinyin for “zhong” with the first tone

4 Improved Gabor-Based Lip Feature Extraction System

Previous research identified that horizontal Gabor features could be used to identify facial features [7]. This resulted in an initial lightweight Gabor-based lip feature extraction system in previous work by the authors [1]. Here, we present an improved visual feature extraction system. Due to space limitations, we will give a general introduction to the system, while focusing on changes. Our system is now quicker, implemented in Python, more accurate, and can calculate height as an additional feature. The feature extraction system is shown in Fig. 3.

Frame Extraction: First, image frames are extracted from a video by using the Python **cap.read()** function.

ROI Identification: Lip regions are extracted using the **Dlib** method, an improvement on the Viola-Jones method used in [1]. We also identify the centre point to select the correct region after Gabor filtering. To demonstrate this, the image in Fig. 4 is labelled with 68 points, with the ROI located using points 6, 10, and 13. The x and y centre co-ordinates are calculated as follows:

$$X = (point(48).x - point(6).x) + \frac{point(54).x - point(48).x}{2} \tag{1}$$

$$\begin{aligned} Y = (point(51).x - point(13).x) + (shape.part(62).y - shape.part(51).y) \\ + \frac{point(66).x - point(62).x}{2} \end{aligned} \tag{2}$$

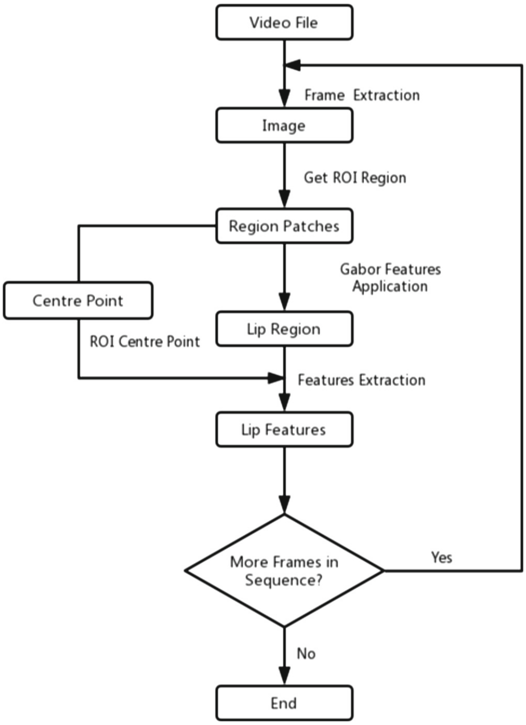


Fig. 3. Key steps of lip feature extraction

Gabor Features: Lip region was identified by using GWT. In OpenCV Python, this can be done using **cv2.getGaborKernel(**$ksize$**,** σ**,** θ**,** λ**,** γ**,** ψ**,** $ktype$**)**. This requires a number of parameters, and after experimentation, optimal parameters for this dataset are listed in Table 1.

The parameters chosen in Table 1 are heavily dependent on factors such as image size, distance from the camera, and speaker pose. The parameters were suitable for the majority of videos, although some occasional small adjustments ± 1 were made to optimise results. It should be noted that no changes were needed within video sequences. Example results are shown in Fig. 5. Here, Fig. 5(a) shows an original frame, and (b) and (c) show an extracted open and closed mouth respectively. The effect of the GWT is shown in Fig. 5(d) and (e) showing an open mouth and a closed mouth. It can be seen that the dark area of the open mouth is obvious, but the closed mouth is very faint. Finally, Fig. 5(f) shows precise feature extraction, as will be discussed next.

Feature Extraction: The Python function **skimage.filters.threshold_yen** determines whether each pixel in transformed image should belong to the target or background region, calculated automatically depending on the individual image. This produces a corresponding binary image. This method returned a

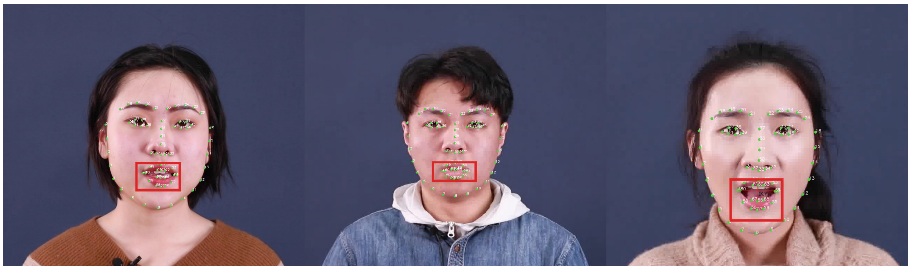
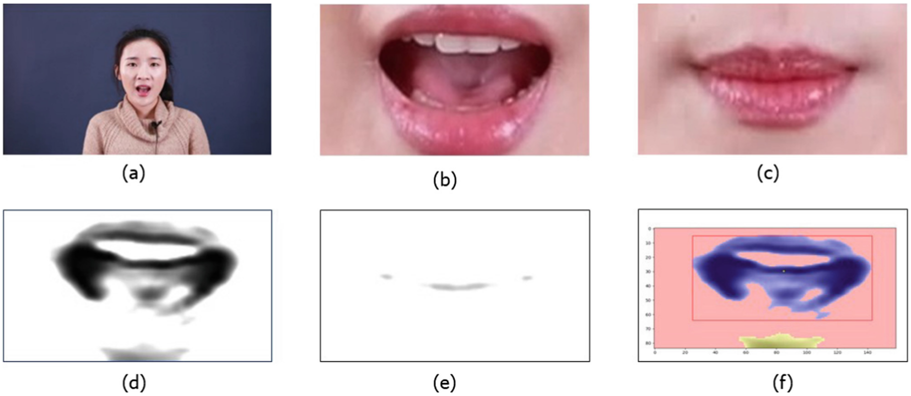


Fig. 4. The ROI and centre points of example speech frames

Table 1. Suitable parameters for Gabor-based feature extraction of AVMC data

Parameter	Description	Value
Wavelength (λ)	Cosine factor of the Gabor filter kernel	16
Orientation (θ)	Orientation of the normal to the parallel stripes of a Gabor function	90
Phase offset (ϕ)	Phase offset of Gabor cosine factor	0
Aspect ratio (γ)	The ellipticity of the support of the Gabor function	0.5
St. deviation (σ)	The standard deviation of the Gabor filter Gaussian function	4
ktype	The type and range of values that each pixel in the Gabor kernel can hold	CV_32F
Ksize	The size of the Gabor kernel	12



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 5. Example of the feature extraction process from one frame

higher threshold value which more accentuated lip features in comparison to the **skimage.filters.threshold_otsu** function.

The Python **regionprops** function is then used to label all connected pixels and to find the mouth region according to the distance between ROI centre point and the centre point of the labelled regions. The **Regionprops** function identifies region properties in an image, and the result of GWT is a number of regions, with the closest region to the ROI centre point being chosen as the lip region R, with several parameters then obtained. Figure 5(f) shows a final result. Seven properties of the blue region represent the lip features. These are:

- Width: The width of R.
- Height: The height of R. Width and height temporal changes provide intuitive and effective lip information.
- Centre (X,Y): Centre point of R. Tracks lip position.
- Mass: Sum of each pixel value within R. Provides 3D mouth depth information, including tongue and teeth changes.
- Area: Number of pixels within R. A clear measurement of mouth openness.
- Orientation: the mouth angle in degrees, mapping pose of speaker.

Table 2. Configuration of Pinyin recognition network for visual speech recognition

Input Layer: (None, 35, 7)					
Conv1D filter: 64 kernel: 3 strides: 1 padding: same activation: linear bias: true	**Conv1D** filter: 128 kernel: 8 strides: 1 padding: same activation: linear bias: true	**Conv1D** filter: 256 kernel: 13 strides: 1 padding: same activation: linear bias: true	**Conv1D** filter: 256 kernel: 18 strides: 1 padding: same activation: linear bias: true	**Conv1D** filter: 256 kernel: 25 strides: 1 padding: same activation: linear bias: true	**Conv1D** filter: 960 kernel: 35 strides: 1 padding: valid activation: linear bias: true
MaxPooling1D strides: 2 pool:2 padding: valid	**MaxPooling1D** strides: 2 pool:2 padding: valid	**MaxPooling1D** strides: 2 pool:2 padding: valid	**MaxPooling1D** strides: 2 pool:2 padding: valid	**MaxPooling1D** strides: 2 pool:2 padding: valid	
Concatenate					
Add					
Batch Normalization					
Conv1D filter: 256 kernel: 6 stride: 1 padding: valid activation: linear bias: true					
MaxPooling1D strides: 2 pool: 2 padding: valid					
Dropout rate: 0.25					
TimeDistributed(Dense) units: 31 activation: softmax bias: true					
Output Layer: (None, 6, 31)					

Table 3. Precision, recall, and f1-score for all Pinyin characters

	precision	recall	f1-score
0	0.743	0.949	0.833
a	0.597	0.573	0.583
b	0.535	0.618	0.573
c	0.559	0.284	0.374
d	0.187	0.186	0.182
e	0.624	0.183	0.281
f	0.067	0.040	0.050
g	0.494	0.099	0.164
h	0.551	0.250	0.343
i	0.541	0.714	0.615
j	0.331	0.311	0.317
k	0.393	0.200	0.257
l	0.316	0.459	0.368
m	0.470	0.435	0.444
n	0.450	0.243	0.313
o	0.511	0.147	0.227
p	0.462	0.250	0.321
q	0.294	0.375	0.328
r	0.579	0.500	0.532
s	0.324	0.486	0.387
t	0.606	0.375	0.457
u	0.550	0.482	0.507
v	0.800	0.200	0.314
w	0.212	0.300	0.247
x	0.343	0.361	0.349
y	0.238	0.391	0.295
z	0.258	0.267	0.261
1	0.431	0.453	0.440
2	0.438	0.192	0.255
3	0.491	0.685	0.569
4	0.512	0.524	0.516
total	0.449	0.372	0.377

4.1 Lipreading with Inception-ResNet, Results and Discussion

To test our improved feature extraction system with our new corpus, we perform visual speech recognition in an Inception-ResNet network [16]. This combines the Inception CNN module with residual connections. The Inception module has an efficient utilization of the computing resources inside the network, which increases network depth and width while keeping the computational cost low. Residual connections have been shown to contribute to faster training of the Inception network [15]. Based on initial trials, the proposed architecture of this network is shown in Table 2. This model is character based, meaning that rather than estimating the Pinyin, given a sequence of visual vectors, it attempts to identify each character.

After 5 runs, the overall character level validation accuracy of this system is 64.96% with an Interquartile Range (IQR) of 0.003, and the CER is 57.72% with an IQR of 0.005. This is in line with other visual systems that consider only single words without sentence level context. The detailed results are shown in Table 3. However, some analysis of the table is required. Firstly, a single model was trained for all characters, and this meant there were more cases of 0 (i.e. no character present), which is a result of having more smaller Pinyin words (i.e. '1yi' has 3 more '0' characters than '1zhong'). Despite the data being evenly distributed at Pinyin level, at character level, this results in a skew to '0', affecting the results. It should also be noted that when the character results are combined into a single Pinyin character, the overall word accuracy is only 11.76%. However, this is not an unexpected result.

Table 3 shows that according to both accuracy scores and f1-scores of 26 letters, the most accurate characters are 'a', 't', and 'r'. Furthermore, 'v' is distinctive, with highest precision but low recall, which may due to the lower distribution rate. It should also be pointed out that the results can be grouped into initials, finals, shared initials and finals, and tones. Excluding zeros, the respective mean f1-score for the 16 dedicated initials is 0.327, for the 6 dedicated finals it is 0.421, for the 3 shared initials and finals, it is 0.335, and for tones it is 0.445. This suggests that performance is better for finals than initials, with the shared initials and finals likely causing confusion. In addition, the overall validation accuracy is likely to be much higher for finals than for initials, which is unsurprising, given that more information is present in the visualisation of a final. Also, conventional deep learning algorithms are likely to not perform well in this test, due to a lack of training data. So above all, an overall character level validation accuracy of 64.96% is along the lines of what is expected.

The results support that recognition is being performed accurately in many cases, considering that visual information alone will only ever generate incomplete information, and that tone is overwhelmingly a function of the vocal cord, and is not visualised. To generate accurate Pinyin matching, then separate models should be trained for initials, finals, and tones, since we are dealing with three distinct word components. We can also improve the very low word accuracy scores by training these separate models, and also improving the formulation of Pinyin, post character generation. There is always likely to be skew in the data, this is a feature of character-level Pinyin generation, and the structure of Pinyin makes direct comparison to English language lip-reading challenging.

Much research has found that it is possible to train deep learning models at the character level, as it allows the model to learn internal word structures. However, to the best knowledge of the authors, this approach has not been used widely for Chinese, with research limited to been text based work, such as by Huang and Wang [9]. Due to the new approaches, we have used in this paper, and the new form of analysis, there is no direct comparison between our results and other results in the literature, as other research tends to consider the Pinyin at a word level, rather than considering it separately at a character level, which as our research has shown, is an important factor. Overall, we can conclude that firstly, the improved feature extraction method improved on that proposed by Abel et al. [1], using a more advanced ROI detection method (Dlib rather than Viola-Jones). Qualitative comparisons identified that our features were more reliable and accurate, with additional features (i.e. height measurements). In addition, using Gabor features rather than CNN features is more suitable for a smaller corpus, and also reduced the training time considerably. Our new AVMC corpus is available on request, and fills a role of clean Chinese speech data, focusing on individual characters, that other corpora do not currently meet. Finally, the initial experiments presented in this paper suggested that results were along the lines of what was expected, and that using a different approach to what is used for non-tonal languages will improve results.

5 Conclusions

This paper introduced a new audiovisual speech corpus, AVMC, which is recorded in a good quality studio environment, and contains 10 speakers reading distinct Chinese characters from a teleprompter. This dataset is fully labelled, and is freely available on request. To demonstrate its effectiveness, we performed initial visual speech recognition with an improved lightweight Gabor-based feature extraction system, with character based recognition results of 64.96% for recognition rate and 57.72% CER when trained with a deep neural network. Future work will be to improve the feature extraction process by making it more robust and easier to configure, as well as improving our database by adding more speakers and more utterances to make it more suitable for deep learning use, and considering training with individual models for different speech components. In addition, for the recognition model, it could be improved to a different Inception-ResNet network by adding 1*1 convolution to reduce the computational cost [15]. Furthermore, after getting lip Gabor images, the stacked denoising autoencoders need to be used to reduce noisy and hypothesis ROI regions [8].

Acknowledgments. This work was supported by XJTLU Grant RDF 16-01-35, and partially funded by the Research Institute of Big Data Analytics.

References

1. Abel, A., Gao, C., Smith, L., Watt, R., Hussain, A.: Fast lip feature extraction using psychologically motivated Gabor features. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1033–1040. IEEE (2018)
2. Abel, A., Hussain, A.: Novel two-stage audiovisual speech filtering in noisy environments. Cogn. Comput. **6**(2), 200–217 (2014)
3. Abel, A., Hussain, A.: Cognitively Inspired Audiovisual Speech Filtering. SCC, vol. 5. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-13509-0
4. Assael, Y.M., Shillingford, B., Whiteson, S., De Freitas, N.: LipNet: end-to-end sentence-level lipreading (2016)
5. Bhadu, A., Tokas, R., Kumar, D.V.: Facial expression recognition using DCT, Gabor and Wavelet feature extraction techniques. Int. J. Eng. Innovative Technol. **2**(1), 92–95 (2012)
6. Cao, J.: Chinese pronunciation: the complete guide for beginner. https://www.digmandarin.com/chinese-pronunciation-guide.html
7. Dakin, S.C., Watt, R.J.: Biological "bar codes" in human faces. J. Vis. **9**(4), 2.1–10 (2009)
8. Han, J., Zhang, D., Hu, X., Guo, L., Ren, J., Wu, F.: Background prior-based salient object detection via deep reconstruction residual. IEEE Trans. Circ. Syst. Video Technol. **25**(8), 1309–1321 (2014)
9. Huang, W.: Character-level convolutional network for text classification applied to Chinese corpus (2016)
10. Hursig, R.E., Zhang, J.X., Kam, C.: Lip localization algorithm using Gabor filters (2011)
11. Petridis, S., Wang, Y., Li, Z., Pantic, M.: End-to-end multi-view lipreading. In: British Machine Vision Conference, London, September 2017
12. Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
13. Sterpu, G., Harte, N.: Towards lipreading sentences with active appearance models. arXiv preprint arXiv:1805.11688 (2018)

14. Sujatha, B., Santhanam, T.: A novel approach integrating geometric and gabor wavelet approaches to improvise visual lip-reading. Int. J. Soft Comput. **5**, 13–18 (2010)
15. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
16. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
17. Wand, M., Koutník, J., Schmidhuber, J.: Lipreading with long short-term memory. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6115–6119. IEEE (2016)
18. Weng, X.: On the importance of video action recognition for visual lipreading. arXiv preprint arXiv:1903.09616 (2019)
19. Zhang, X., Gong, H., Dai, X., Yang, F., Liu, N., Liu, M.: Understanding pictograph with facial features: end-to-end sentence-level lip reading of Chinese (2019)
20. Zhou, Z., Zhao, G., Hong, X., Pietikäinen, M.: A review of recent advances in visual speech decoding. Image Vis. Comput. **32**(9), 590–605 (2014)