An Automatic Lip-reading Method Based on Polynomial Fitting

Meng Li

Abstract

This paper addresses the problem of speaker-dependent isolate digits recognition using sole visual information. We employ intensity transformation and spatial filter to estimate the minimum enclosing rectangle of mouth in each frame. Thus, for each utterance, we can obtain two vectors composed of width and height of mouth, respectively. Then, we propose an approach to recognize the speech based on polynomial fitting. Firstly, both width and height vectors are normalized into the constant length via interpolation. Secondly, least square method is utilized to produce two 3-order polynomials that can represent the main trend of the two vectors, respectively, and reduce the noise caused by the estimate error. Lastly, positions of three crucial points (i.e. maximum, minimum and right boundary point) in each 3-order polynomial curve are recorded as a feature vector. For each utterance, we calculate the average of all vectors of training sample to make a template, and using Euclidean distance between the template and testing data to perform the classification. Experiments show the promising results of the proposed approach.

1 Introduction

Lip reading is to understand speech by visually interpret the lip movement of speakers [7]. This technique has potential attractive applications in speech recognition, human identification, and so forth [4, 9, 13].

So far, two kinds of features are widely used in lip reading system, namely image-based and model-based. In the image-based approach, the pixels of lip region are transformed by PCA, DWT or DCT, to become a feature vector [2, 5, 6]. Under the ideal environment, the accuracy of image-based approach is considerable high, but the performance will be degraded seriously in real environment. One main reason is that the approach is restricted by the illumination, mouth rotation and some other conditions. Thus, from the practical view point, the image-based approach is not the appropriate choice for automatic lip reading system. In the model based approach, shape and position of lip contours, tongue, teeth or some other features like width and height of mouth are modeled, and controlled by a set of parameters(e.g. Snake, ASM and AAM) [3, 8, 10, 12]. Model-based approach can invariant to the effects of scaling, rotation, translation and illumination. Hereby, with regard to the feature extraction, many researchers resort to the modal-based approach.

In this paper, we propose a lip reading approach under simple modal – the width and height of mouth. We employ intensity transformation and spatial filter to the image of ROI(Range Of Interesting) in gray scale space to localize the minimum enclosing rectangle of lip automatically. Then, give a video clip for an utterance, we can obtain two vectors composed of the width and hight of mouth from each frame, respectively. Based on the least square method, two 3-order polynomials are built to fit the width and height vector. The positions of three crucial points (i.e. maximum, minimum, and right boundary point) in each 3-order polynomial curve are recorded as a feature vector. For each utterance, we calculate the average of all vectors of training data to make a template, and use Euclidean distance between the template and testing data to perform the classification.

The remainder of this paper is organized as follows. Section 2 describes the visual feature, namely the minimum enclosing rectangle of mouth, extraction method. Section 3 presents our new approach for lip-reading recognition. In Section 4, we will conduct the experiment to compare our approaches with the existing methods. Finally, Section 5 draws a conclusion.

2 Lip localization and feature extraction

Before showing the details, a pre-processing is needed. The images captured by camera are comprised of RGB values. We heuristically project these RGB values into the gray-level space based on the following equation:

$$I = 0.299R + 0.587G + 0.114B. \quad (1)$$

In order to enhance the contrast between lip and surrounding skip region, we adjust the histogram of the image and make it equalized for the first step. Then we make an

accumulation of gray level value for each row of the image. The slopes of the curve contain the information about the boundaries between the lips and the surrounding skin region. The minimum value on the curve retained as the row position of mouth corner points or the nearby position, the row can be named as horizontal midline of mouth. The midline is shown in Figure 1.

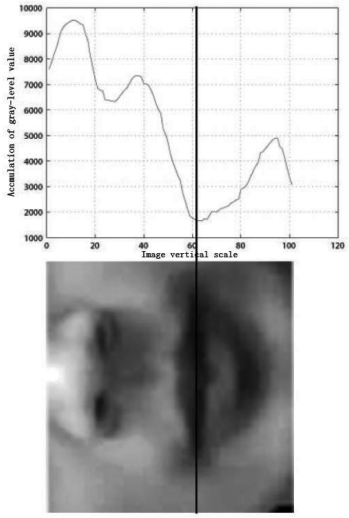


Figure 1. Accumulation curve of gray level value for each row. The vertical crossing line represents the relation between the horizontal midline of mouth and the minimum value of the accumulation curve.

The curve of gray level values along with the horizontal midline is saved in vector G. Building a sub-vector G_s by a segment of G which between the first maximum from left and the first maximum from right. Using the following equations to make the curve smooth and save it into a new vector which named C.

$$C_l^{(i)} = \begin{cases} G_s^{(i)} & (C_l^{(i-1)} > G_s^{(i)}) \\ C_l^{(i-1)} & (C_l^{(i-1)} \leq G_s^{(i)}) \end{cases} \quad i = 1, 2, \ldots, n \quad (2)$$

$$C_r^{(i)} = \begin{cases} G_s^{(i)} & (C_r^{(i+1)} > G_s^{(i)}) \\ C_r^{(i+1)} & (C_l^{(i+1)} \leq G_s^{(i)}) \end{cases} \quad i = n-1, n-2, \ldots, 1 \quad (3)$$

$$C = C_l + C_r \quad (4)$$

where C_l and C_r are assistant vectors, $C_l^{(i)}$ is the ith element in vector C_l, n is the dimension of the vector G. The initial values of the two vectors are shown in equation 5.

$$\begin{aligned} C_l^{(1)} &= G^{(1)} \\ C_r^{(n)} &= G^{(n)} \end{aligned} \quad (5)$$

Set the minimum of the most left and most right value in C as threshold. Elements in C less than the threshold build a new vector C' . Accordingly, the average can be calculated by the equation 6.

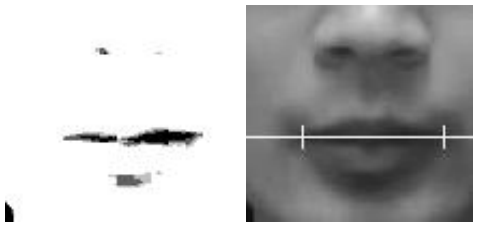
$$c_{avg} = \frac{\sum_{i=1}^{m} C'^{(i)}}{m} \quad (6)$$

The equation 7 is employed to adjust the contrast of image.

$$I_{out} = \begin{cases} 255 & (1.5c_{avg} < I_{in} < 1) \\ \frac{500}{c_{avg}} - 500 & (0 < I_{in} \leq 1.5c_{avg}) \end{cases} \quad (7)$$

where I_{in} is the input gray level value, and the I_{out} is the output.

For the adjusted image, a 11×1 searching block is performed along with the midline, the positions of the most left and right non-all white block are marked as the column of mouth corner candidates. The procedure is performed through an iterative process in the steps above, repeated until the position of mouth corner candidates no longer changed or the image turned into binary. Figure 2 illustrate the result of contrast adjustment and the the corresponding mouth corner estimate result.



(a) (b)

Figure 2. Image of contrast adjust result (a) and mouth corner estimate result (b).

As shown in equation 8 and 9 , a 3×3 mask is employed to perform mean filter in the initial image.

$$M = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{bmatrix} \quad (8)$$

$$I^{(i+1)} = I^{(i)} * M \quad (9)$$

where the $I^{(i)}$ is the result of ith time filter. The times filter performed is determined by the equation 10.

$$\delta_i = dist(I^{(i+1)}, I^{(i)}) \tag{10}$$

where δ_i is the Euclidean distance between $I^{(i)}$ and $I^{(i+1)}$. The procedure should be ceased once δ_i less than a given threshold, and the $I^{(i+1)}$ can be marked as I_f.

Due to the position of left and right mouth corners have been estimated in section 2.1, we can utilize them to calculate the center of mouth easily. For each $I^{(i)}$, a gray value vector $G_{mu}^{(i)}$ is built by the segment from the center point to the top of image along with the normal direction respectively. Then the vector ΔG_{acc} is calculated by the equation 11.

$$\Delta G_{acc} = \sum_{i=1}^{n} (|G_{mu}^{(0)} - G_{mu}^{(i)}|) \tag{11}$$

The point correspond to extreme value of maximum (except boundary value) is retained as the row position of upper bound of mouth.

Then the subtracted image between $I^{(0)}$ and I_f can be calculated. For observing conveniently, an image inversing transformation is employed. The result is shown in Figure 3.

Figure 3. The subtracted image between source gray-scale $I^{(0)}$ and the filtered image I_f. For observing conveniently, an image inversing transformation is employed.

We get the gray level value along with the normal direction pass the middle point of mouth to the bottom of the image. The point perform extreme value of minimum (except boundary value) is retained as the row position of lower bound of mouth. The estimate of mouth upper and lower bound is shown in Figure 4.

3 Recognition method

For one utterance procedure, e.g. a isolate digit or a word, we can capture the video clip of speaker's lip motion, and split it into a frame sequence. Then, utilize the method

Figure 4. The estimate of mouth upper and lower bound.

described in Section 2 in each frame, we can get two vectors composed of width and height of mouth respectively. Since the time for each pronunciation is inconstant, a interpolation method is employed to make the length of two vectors same (in our paper, the length is 100), this two vectors are marked as F_w and F_h. A example is shown in Figure 5.

Since the range of lip motion for different people is inconstant, we use the radio between displacement and the original position of lip to represent the trend of motion. The normalization method is shown in equation 12 and 13.

$$F_w^{norm} = K\frac{(F_w - F_{w_1})}{F_{w_1}} \tag{12}$$

$$F_h^{norm} = K\frac{(F_h - F_{h_1})}{F_{h_1}} \tag{13}$$

where F_{w_1} is the first element in F_w, and K is a gain coefficient which make the motion trend more significant, in this paper, the value of K we get experimently is 30.

Then, we employ least square method so as to find two polynomials to fit F_w^{norm} and F_h^{norm}. The polynomial is like equation 14:

$$P = \sum_{k=0}^{n} a_k x^k \tag{14}$$

To find the coefficient a_k, we should make the following equation minimum.

$$I = \sum_{i=0}^{m} (\sum_{k=0}^{n} a_k x_i^k - y_i)^2 \tag{15}$$

The minimum of I is found by setting the gradient to zero. Since the equation 15 contains m parameters there are m gradient equations.

$$\frac{\partial I}{\partial a_i} = 2\sum_{i=0}^{m} (\sum_{k=0}^{n} a_k x_i^k - y_i) x_i^k = 0 \tag{16}$$

where y_i is the $F_{w_i}^{norm}$ or $F_{h_i}^{norm}$, m is the maximum index of vector which equal to 99, and x_i equal to $0.1i$ so as to

avoid the conditioned in coefficient matrix. Moreover, owing to the characteristic of human speech, n is chosen by 3. Thus, we can get the solution $A_w = [a_{w0}, a_{w1}, a_{w2}, a_{w3}]^T$ and $A_h = [a_{h0}, a_{h1}, a_{h2}, a_{h3}]^T$. A example of polynomial fitting result is shown in Figure 6.

The polynomial shapes for the same utterance are constrained to have similar expression. Thus, we can get the global maximum and minimum in the two polynomial marked as $(x_{w_{min}}, y_{w_{min}})$, $(x_{h_{min}}, y_{h_{min}})$, $(x_{w_{max}}, y_{w_{max}})$ and $(x_{h_{max}}, y_{w_{max}})$ to build the feature vectors which shown below:

$$F_w = [x_{w_{min}}, y_{w_{min}}, x_{w_{max}}, y_{w_{max}}, y_{w_{bound}}]^T \quad (17)$$

$$F_h = [x_{h_{min}}, y_{h_{min}}, x_{h_{max}}, y_{h_{max}}, y_{h_{bound}}]^T \quad (18)$$

where the y_{bound} is the most right value of polynomial when $x \in [0, 9.9]$ (e.g. $x = x_{99}$).

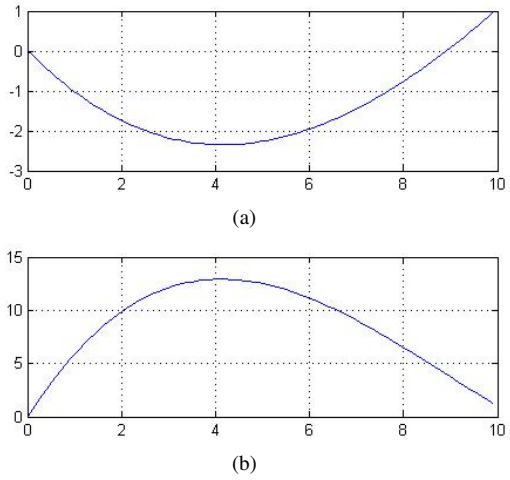


Figure 6. The curve of polynomial which fit to the width vector (a) and height vector (b) shown in Figure 5.

Moreover, for the feature vectors, if $x_{min} \notin [0, 9.9]$, both x_{min} and y_{min} are set to zero. In a similar way, we can determine the value of x_{max} and y_{max}. For example, as shown in Figure 6, the feature vectors are $F_w = [4.1558, -2.3476, 0, 0, 0.9460]^T$ and $F_h = [0, 0, 4.1266, 12.9164, 1.2969]^T$.

For each utterance, we calculate the average of all vectors of training data to make a template by following equations.

$$T_w^{i+1} = \frac{T_w^i + F_w}{2} \quad (19)$$

$$T_h^{i+1} = \frac{T_h^i + F_h}{2} \quad (20)$$

Digit	Phonetic Symbol	Digit	Phonetic Symbol
0	[liŋ]	5	[uː]
1	[jau]	6	[liou]
2	[ɤɻ]	7	[tɕʻ]
3	[san]	8	[baː]
4	[sɿ]	9	[tɕiou]

Table 1. The pronunciations of number 0 to 9 in Chinese mandarin.

where Ti is the template after i times trains employed, F is the new training data. For each classification, there is a template which involve two vectors.

When testing, we calculate the distance between each template and testing data via the following equation.

$$d = ||F_w - T_w|| \cdot ||F_h - T_h|| \quad (21)$$

that is, the input testing data is classified into the category which corresponding into the minimum d.

4 Experiment result

We conduct an experiment to demonstrate the performance of the proposed approach. The experiment environment is shown in Figure 7. The illumination source is a 18w fluorescent lamp which placed in front of speaker.The resolution of camera is 320×240, and the FPS(Frames Per Second) is 30.

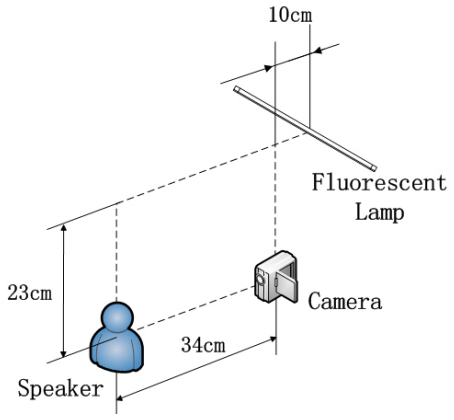
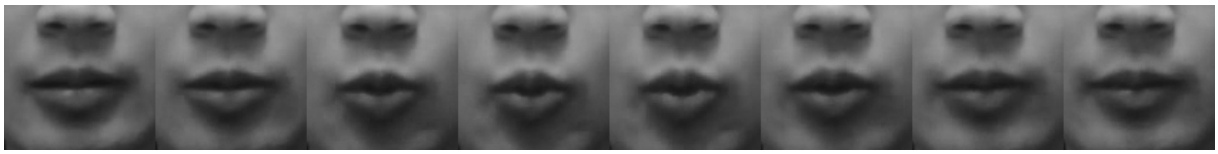


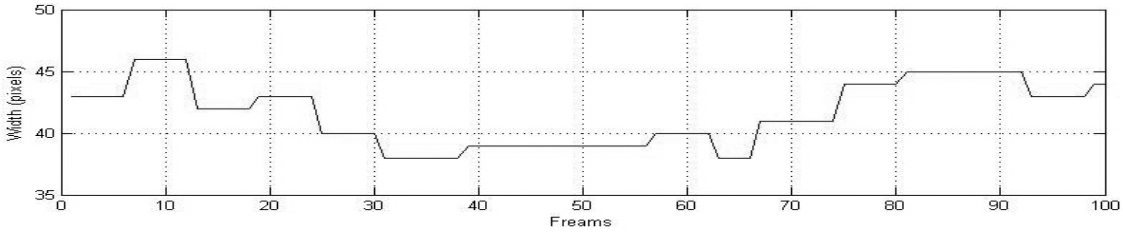
Figure 7. The illustration of experiment environment.

Our task is to recognize 10 isolate digits(0 to 9) in Chinese mandarin, which pronunciations are shown in Table 1.

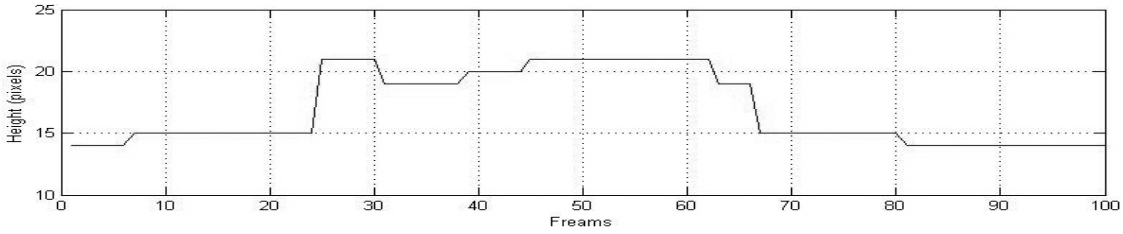
There are 5 speakers(4 males and 1 female) take part into the experiment. For each digit, speakers were asked to re-



(a)



(b)



(c)

Figure 5. Some frames of the pronunciation of "5" in Chinese mandarin (a), and the corresponding width vector (b), height vector (c). In this illumination, although there are some noise caused by the estimate error, we can see that the main trend of width is original-narrow-original, and the height is original-high-original.

Digit	Accuracy	Digit	Accuracy
0	0.972	5	0.912
1	0.952	6	0.964
2	0.976	7	0.744
3	0.964	8	0.952
4	0.788	9	0.932

Table 2. The pronunciations of digits 0 to 9 in Chinese mandarin.

peat ten times to train the system, and fifty times to test. Figure 8 shows the performance of our approach with the different number of training samples. Moreover, Table 2 shows the recognition accuracy for each number with 10 training samples.

In order to compare with some existed approaches, we find two paper which also using width and height of mouth as the visual feature to perform the lip reading recognition. In [11], the recognition accuracy achieved under three

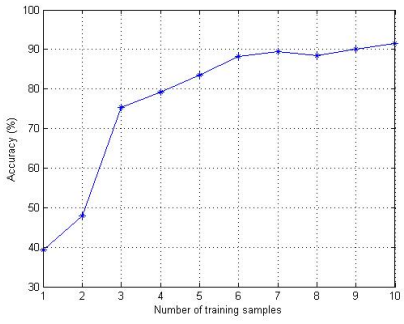


Figure 8. The testing result in different number of training samples.

Method	Accuracy
HMM	81.27%
RDA	77.41%
Spline	91.49%
ST Coding	77.20%
Our approach	91.56%

Table 3. The lip reading recognition accuracy obtain using different approaches.

	0	1	2	3	4	5	6	7	8	9
4	-	-	-	4	197	-	7	42	-	-
7	-	-	-	1	63	-	-	186	-	-

Table 4. The classification detail of "4" and "7" when number of training samples is 10.

method(HMM based, RDA based, and Spline based); the task is to recognize the isolate digits 0 to 9 in English. Moreover, in [1], the classification method is ST(spatio-temporal) coding based; the task is to recognize the isolate digits 0 to 9 in French. Table 3 shows the recognition performance obtained using different approaches.

Nevertheless, we have found that it is not enough to utilize the visual feature, i.e. the width and height of mouth, to distinguish some Chinese utterance, e.g. "4" and "7". Table 4 shows the classification detail of the two digits when number of training samples is 10. In future work, we will focus on how to find a more appropriate modal of mouth for the purpose of lip reading.

5 Conclusion

In this paper, we have proposed a new approach to automatic lip reading recognition based upon polynomial fitting. The feature vector of our approach have low dimensions and the approach need small testing data set. Experiments have shown the promising result of the proposed approach in comparison with the existing methods.

References

[1] A.R.Baig, R.Séguier, and G.Vaucher. Image sequence analysis using a spatio-temporal coding for automatic lip-reading. In *Proc. IEEE International Conference on Image Analysis and Processing*, pages 544–549, Venice, Italy, 1999.

[2] C.Bregler and Y.Conig. "eigenlips" for robust speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech, Signal Processing*, pages 669–672, Adelaide, Australia, 1994.

[3] C.Neti, G.Iyengar, G.Potamianos, A.Senior, and B. Maison. Perceptual interfaces for information interaction: Joint processing of audio and visual information for human-computer interaction. In *Proc. International Conference on Spoken Language Processing*, Beijing, China.

[4] G.Potamianos, C.Neti, J.Luettin, and I.Matthews. Audio-visual automatic speech recognition: An overview. In G.Bailly, E.Vatikiotis-Bateson, and P.Perrier, editors, *Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004.

[5] G.Potamianos and H.P.Graf. Discriminative training of hmm stream exponents for audio-visual speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3733–3736, Seattle,WA, 1998.

[6] G.Potamianos, J.Luettin, and C.Neti. Hierarchical discriminant features for audio-visual lvcsr. In *Proc. IEEE Internatioal Conference on Acoustics, Speech, Signal Processing*, pages 165–168, Salt Lake City, Utah, USA, 2001.

[7] J.Bulwer. *Philocopus, or the Deaf and Dumbe Mans Friend*. Humphrey and Moseley, 1648.

[8] J.Luettin and N.A.Thacker. Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178, 1997.

[9] J.Luettin1, N.A.Thacker, and S. Beet. Speaker identification by lipreading. In *Proc. IEEE International Conference on Spoken Language Processing*, pages 62–65, Philadelphia,USA, 1996.

[10] S.Dupont and J.Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151, 2000.

[11] S.L.Wang, W.H.Lau, A.W.C.Liew, and S.H.Leung. Automatic lipreading with limited training data. In *Proc. IEEE International Conference on Pattearn Recognition*, pages 881–884, 2006.

[12] S.Werda, W.Mahdi, and A.B.Hamadou. Colour and geometric based model for lip localisation: Application for lip-reading system. In *Proc. IEEE International Conference on Image Analysis and Processing*, pages 9–14, Modena, Italy, 2007.

[13] T.Chen and R.R.Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–851, 1998.