

Lip-Reading: Toward Phoneme Recognition Through Lip Kinematics

Ak Muhammad Rahimi Pg Hj Zahari

Abstract Heuristic parameters such as width and height are usually obtained in audio-visual speech recognition. However, the presence of noise has an impact on such system. In the paper, we present a mathematical study investigating whether descriptive parameters derived from lip shapes can improve the performance of the system through the use of a mathematical model. The video database used consists of five separate pronunciations of the numbers ranging from 0 to 9. Three categories of data have been successfully classified; the polynomial coefficient (curving of the lips), width and height (both inner and outer) and also the raw data (coordinates). The results showed that the best classifier is the curving of the bottom lip contour with an accuracy of 90.91% and the weakest classifier is from points on the right upper lip contour with accuracy of 12.24%.

Keywords Noise · Mathematical model · Polynomial coefficients · Classifier

1 Introduction

In the presence of audible noise, the performance of speech recognition systems becomes degraded. One method of reducing the effects of noise in such systems is to make use of visual information that can be obtained from the speaker and in particular the movements of the lips. Lip reading is therefore seen as a supporting process to speech recognition where its application in stand-alone process ranges from the use of mobiles phones in health application to video surveillance use for security. However detecting the lips has become more and more challenging, because of large difference between people in shapes, existence of facial hair, head movement and lighting. To be able to make use of the visual information, features derived from the lip should be extracted and therefore lip models needs to be built.

Ak.M.R.Pg.Hj. Zahari(✉)

Institut Teknologi Brunei, Jalan Tungku Link, Gadong, Brunei Darussalam
e-mail: rahimi.zahari@itb.edu.bn

© Springer International Publishing Switzerland 2016

K. Lavangananda et al. (eds.), *Intelligent and Evolutionary Systems*,
Proceedings in Adaptation, Learning and Optimization 5,

DOI: 10.1007/978-3-319-27000-5_33

In this paper, we proposed a mathematical model of the lips that will be developed by obtaining measurements to estimate both the static parameters that are peculiar to individual speakers and the dynamic changes in these parameters that occur when specified words are uttered. Once a database of parameters has been developed, the models will be used to identify words spoken by further speakers and the results compared with existing approaches in the literature. A publicly available corpus of speakers will be used as well as a new high-definition corpus that is currently being established at Loughborough University. Although mathematical models of lips have been developed by previous researchers (and these may be re-used in this study), there appears to be no previous use of such models for audio-visual speech recognition.

The text is organized as follows: In Sect. 2, previous related works are studied as well as the introduction of the proposed method with classification techniques and in Sect. 3, discussion of the results. Lastly, conclusion of this work will be addressed with further suggestion in Sect. 4.

2 Lip Kinematics

In noisy environments, humans are able to reduce speech recognitions errors by using the speaker's lip movements and indeed many people with hearing difficulties rely on lip reading to provide majority of the speech information they receive. Most of the recent methods for extracting lip contours are based on image segmentation and color-based information of the lip region. However, a lip template can be used to describe lip contour and several curves and special points are employed to approximate actual lip shape in order to obtain geometric feature of the lip shape. Various lip models may be found in the literature. In [1], the lip is made up of two contours, outer and inner. Both of them are described by curves using nonlinear least square methods. In [2], combination of two semi-ellipses was proposed with the employment of 16-point geometric deformable model and initialization of evolving curves. In [3], the introduction of parameterized key points has proven to be the most important aspects for lip movement recognition. These points will help to calculate the height, width and area. Lastly in [4], multiple points' representation of the lip was introduced and the classifications of the words were implemented using Euclidean Distance.

2.1 *The Proposed Method*

The proposed method consists of three different phases with the use of a selected video database. The schematic depicted in Fig.1. Mainly, first phase deals with the extraction of parameter; second phase concerns with classification techniques and the last phase discusses and evaluates the classification results.

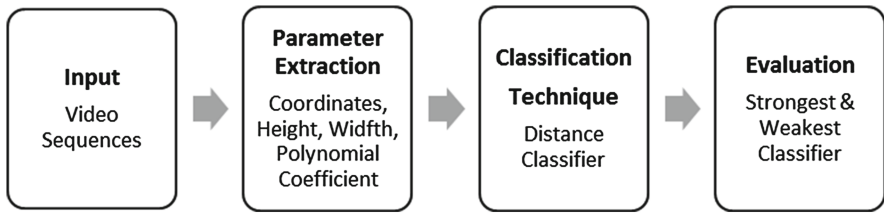


Fig. 1 Schematic Diagram of the proposed method

2.2 Video Images Database

In this paper, the video sequences are obtained from [5]. This has been made using a high definition camcorder. Selection of video was based on the success detection rate of the mouth region. Moreover, it consists of pronunciation of numbers ranging from '0' to '9' and each has 5 different sets labelled as 'a, b, c, d and e'. These sets will be used as comparison means.

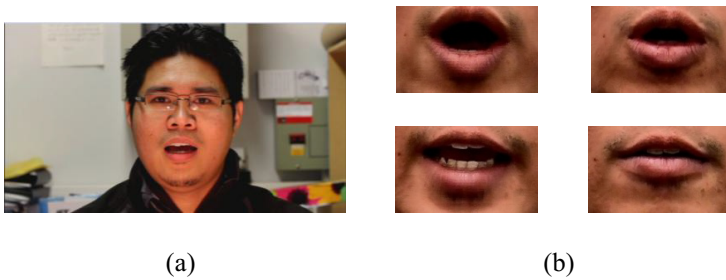


Fig. 2 Sample of video database (a) Face Region (b) Mouth region

2.3 Parameter Extraction

In extracting the parameter, we proposed the idea of manually getting the outline. This implies that the user needs to choose the right point on the image. The implementation of the automatic method has shown its weaknesses in getting the right contour. For the method, specific functions have been made which corresponds to the name containing built-in MATLAB functions which perform image processing tasks. Coordinates from every frame and video will have to be selected and these videos will have a slightly different amount of frames. Manual Identification of the coordinates via user selection takes around 15-30 minutes for each video as each frame needs to be processed, and this also includes the processing time for the arrangement of the result into a database. Through the study of the mathematical model in the literature, we implemented a 21 point model as seen in Fig. 3 and the coordinates are in Table 1.

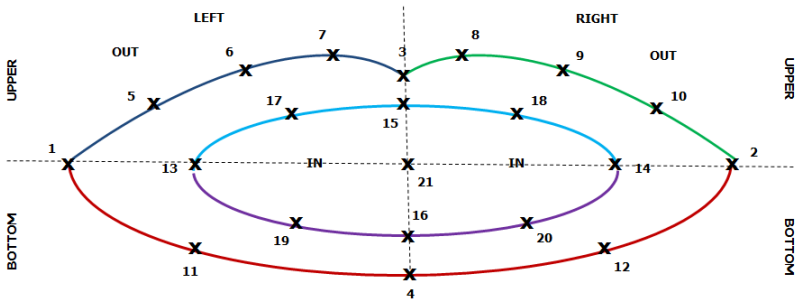


Fig. 3 21-point geometric proposed model

Table 1 Proposed coordinates

1	Outleft	8	UpperOutRight	15	UpperInMid
2	OutRight	9	UpperOutRight1	16	BottomInMid
3	UpperOutMid	10	UpperOutRight2	17	UpperInLeft
4	BottomOutMid	11	BottomOutLeft	18	UpperInRight
5	UpperOutLeft	12	BottomOutRight	19	BottomInLeft
6	UpperOutLeft1	13	InLeft	20	BottomInRight
7	UpperOutLeft2	14	InRight	21	Mid

These coordinates were used to identify five more additional parameters; Outer and Inner height, Outer and Inner width and lastly, polynomial coefficient. Both height and width are calculated using Euclidean Distance based on Eq.1. Suppose we have two coordinates (x_1, y_1) and (x_2, y_2) , the distance, D is as follows:

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{1}$$

In order to get the polynomial coefficient, we implemented a ‘least squares’ method [6]. Given the coordinates, this method will minimize the squared error between the set of measured data and the curve. We then use the result to compute a 2 degree polynomial, i.e. quadratic polynomial. The order of polynomial relates to the number of turning points that can be accommodated. In this case, we eventually come across a turning point.

2.4 Classification Techniques

The obtained data will have some sort of patterns between each other. It is very important to develop proper methodologies to organize them. The idea behind this classification is to assign a class to an unknown or unknown pattern based on previously acquired knowledge about the objects and the classes to which they

belong. However, designing such a pattern recognition system is usually an interactive process that involves the selection and computation of features from the objects that needs to be classified and the numerical data, for instance collections of feature vectors often necessary needs to undergo pre-processing before they can be inputs to any classifier. We have chosen a simple method known as minimum distance for the classification [7]. Assumptions were made to which each time when the test data set is applied to the training data it will give minimum two results. Prior knowledge suggests each of the data used has two correct sets. The process will remove any presence of duplication. As each video has five different sets, it can be divided as three training sets and two tests sets.

The classes that have been considered are divided into three categories:-

- Polynomial coefficient
- Width and Height
- Raw Data (or, Coordinates)

In this paper, the only preprocessing technique being applied is normalization method based on Eq. 2 which will result in the number of frames ranging from 0 to 1. This allows easier comparison of results between each of the videos. Suppose that we have a certain number of frames in a video, i . The variable which contains these values can be represented as f . Thus,

$$NormalizedValue(f_i) = \frac{f_i - f_{\min}}{f_{\max} - f_{\min}} \quad (2)$$

Where

f_{\max} - Maximum number of frames

f_{\min} - Minimum number of frames

Other than normalization, the obtained data is assumed to be correct, and this can be checked in the classification process. The proposed classification is illustrated in Fig. 4.

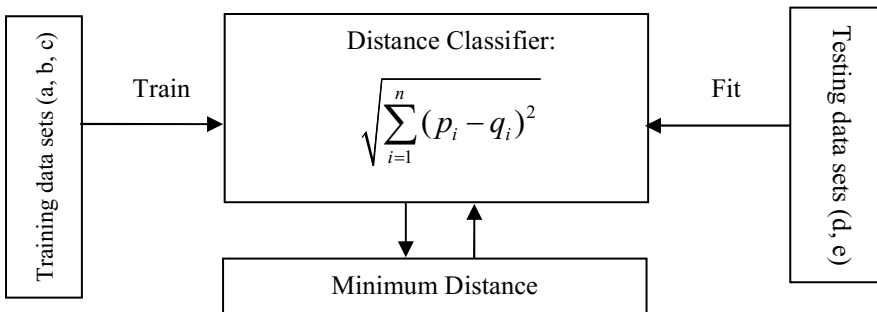


Fig. 4 Chosen classification method

2.5 Evaluation

To evaluate the results of the classification, we studied only the proportion of the total number of predictions that were correct through the use of confusion matrix [8]. The motivation behind this is to check which parameter has a significant change in the pronunciation of numbers and thus can be used to improve the accuracy of detection. The results will undergo two phases shown in Table 2: Phase I is aimed at whether it can be classified correctly to its respective data, basically to classify data correctly and Phase II suggests without the existence of a specific data set will it be able to classify that data from the rest of the data or successfully classify incorrect data. The classifier also can be categorized into two; the strongest classifier and the weakest classifier. The strongest classifier can be found by searching the maximum result of accuracy in the Phase I and also a minimum accuracy in Phase II. The reason behind the maximum suggests that the classifier manage to classify the data correctly whereas the minimum in Phase II shows that the classifier can classify unwanted data. However the weakest classifier will have the opposite characteristics; a minimum in Phase I and a maximum in Phase II.

Table 2 Proposed Coordinates

Phase	Training Sets (a, b, c)	Phase	Training Sets (a, b, c)	Test Sets (d, e)
I	0	II	1 till 9	0-9
	1		0,2-9	
	2		0-1,3-9	
	3		0-2,4-9	
	4		0-3,5-9	
	5		0-4,6-9	
	6		0-5,7-9	
	7		0-6,8-9	
	8		0-7,9	
	9		0 till 8	

3 Results

3.1 Behavior of the Lip

In dealing with such model, basic definition and terminology are quite useful. One of them is the coordinate system which for an image in MATLAB is different from a normal graph. The y-axis is reversed and the axes started from 1 instead of 0. Moreover, units used are in unit pixel. Video '1a' was chosen as an example to illustrate the movement of the lip uttering the number '1'. The video consists of 13 frames altogether. The analysis consists of three parts, the changes in the width, changes in height and as well as the changes in the polynomial coefficients.



Fig. 5 Pronunciation of number '1' (top left to bottom right)

3.2 Changes in Parameter

The behavior of the width and height are seen in Fig. 6. Results suggest that both of the width decrease to a minimum and then slowly increase to a steady state. This implies that the mouth is in the state of protrude. In other words, the lip extends forward which gives pressure in pronouncing 'wu'. And later the increase is when the lip stretches out, uttering 'unn'.

On the other hand, the height behaves oppositely where the height increases to a maximum; this is the effect after the mouth utters 'wu'. It then returns to its original state.

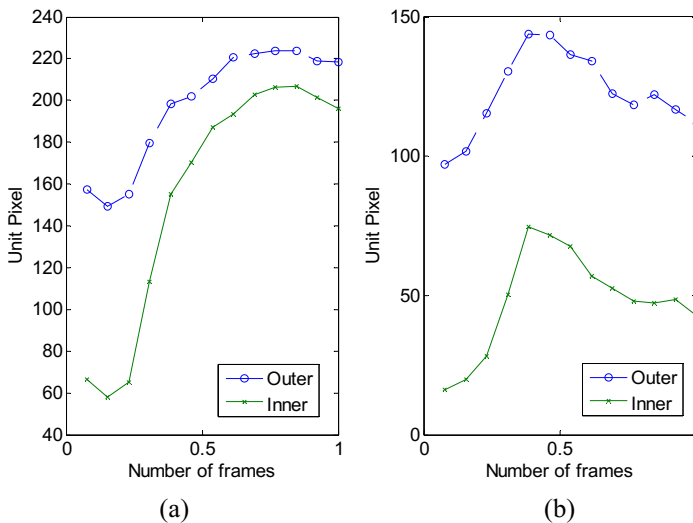
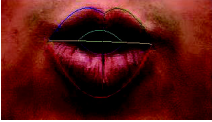
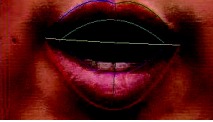
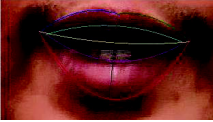
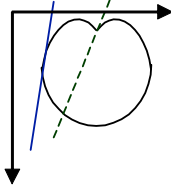
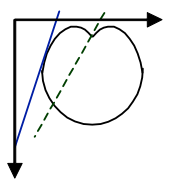
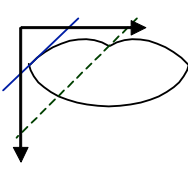
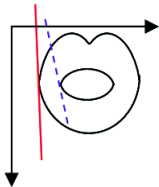
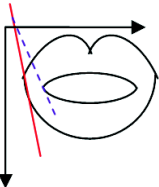
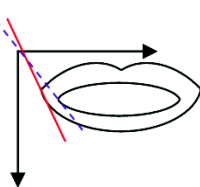
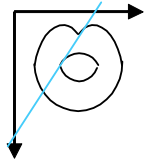
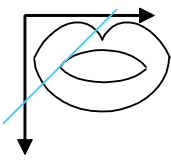
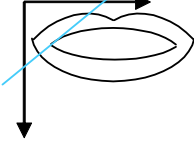


Fig. 6 Results of the movement of the lips (a) Changes in width (b) Changes in height

Aside from the two parameters, the effect of polynomial coefficients was investigated where only the y-intercept has been analyzed. This is due to visible large change in values. Curve representations and observation are illustrated in Table 3.

Table 3 Changes in the polynomial coefficient based on the respective curve

Video Sequence			
Curve 1 (solid line) Curve 2 (dashed line)			
Curve 3 (solid line) Curve 5 (dashed line)			
Curve 4 (solid line)			

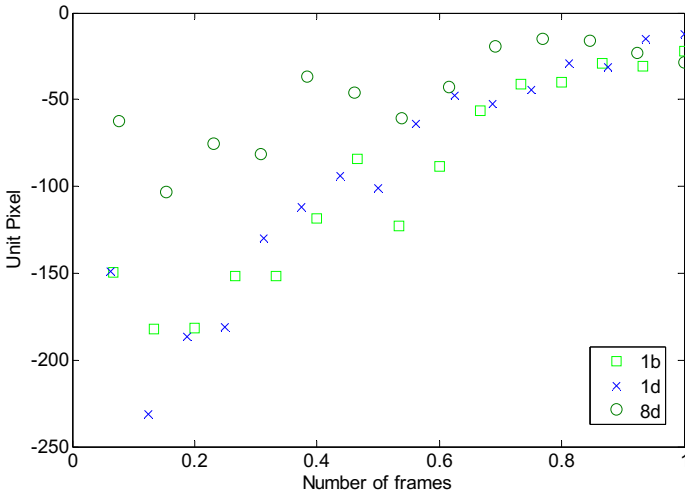
3.3 Discussion

The summary of the classification results based on only the strongest and weakest classifier from each of the three categories is shown in Table 4.

Curve 3, representing the bottom outer lip contour was found to be strongest classifier because of the highest percentage of accuracy compared to the other classifier within all of categories. This implies that each of the data sets has its own unique behavior for this specific coefficient, illustrated in Fig. 7. An example of the training dataset, i.e. 1b is plotted together with both test sets 8d and 1d. Although it seems that they behave in a similar way, the Euclidean distance between 1b and 1d has proven to be smaller than of 8d. Thus it has been successfully classify the data. Other datasets are not shown due to obscure result.

Table 4 Summary of the classification results

	Classifier	Accuracy (%)	
		Phase I	Phase II
Strongest ↓	Curve 3	90.91	1.10
	Inner Height	83.33	1.64
	YBottomInMid	72.73	3.61
	XInLeft	72.00	2.84
	Outer Height	67.86	5.00
	Curve 2	64.29	4.49
	YUpperOutRight	20.00	9.52
Weakest	XUpperOutRight	12.24	10.27

**Fig. 7** Comparing of data resulting from the pronunciation of '1' and '8'

Constant movements of the lips up and down have major effect on the width and height. Results suggest that coordinate InLeft has a high accuracy, especially in the x-direction. This statement is supported by the coordinate which define the width (marked as 'o' in Fig. 8a). InRight, OutLeft and OutRight coordinates have high accuracy as well. The movement in y-direction however, relates to the impact is caused by the changed in height. The Y-Coordinate for UpperInMid and UpperOutMid (marked as 'o' in Fig. 8b) have lower accuracy compared to them as when the mouth is moving, only the bottom part of the lips played a major role. UpperOutRight coordinates, which are connected to the curving of the upper lip is the weakest classifier within the three categories and one factor might be because of the orientation of the image.

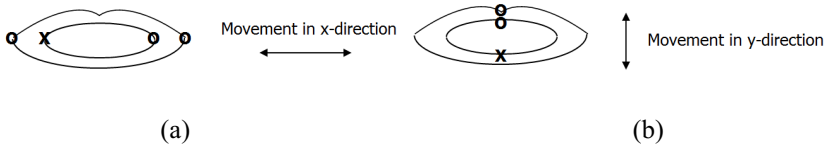


Fig. 8 Illustrations of the lip (a) coordinate InLeft (b) coordinate BottomInMid

4 Conclusion

This paper presented a mathematical study of lip-reading parameters. We successfully implemented a method for extracting such parameter allowing them to be analyzed and classified depending on the training and test data sets used. The exact coordinates are located and tracked depending on the user selection. The system has been tested on the database that contains pronunciations of ‘0’ to ‘9’. We have also discussed the result of the classification and the result has shown that the bottom outer lip contour is the strongest classifier.

Simple functions as well as the understanding of the different variations of behaviours of the lips are discussed throughout the paper which can aid in the distinguishing different between uttering of words. As future work, we intend to explore a higher degree of polynomial for the curve fitting, better classification methods as well as implementation of 3-Dimensional model.

References

1. Liu, H.: Study on lipreading recognition based on computer vision. In: Proceedings of the 2nd International Conference on Information Engineering and Computer Science (2010)
2. Liu, X., Cheung, Y.: A robust lip tracking algorithm using localized color active contours and deformable models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1197–1200 (2011)
3. ur Rehman Butt, W., Lombardi, L.: A survey of automatic lip reading approaches. In: Proceedings of the Eighth International Conference on Digital Information Management (ICDIM 2013), pp. 299–302 (2013)
4. Yargic, A., Dogan, M.: A lip reading application on MS Kinect camera. In: IEEE International Symposium on Innovations in Intelligent Systems and Applications, IEEE INISTA, pp. 1–5 (2013)
5. Ibrahim, M.Z.: A novel lip geometry approach for audio-visual speech recognition (2014)
6. Chi, E.C., Scott, D.W.: Robust Parametric Classification and Variable Selection by a Minimum Distance Criterion. *Journal of Computational and Graphical Statistics* **23**, 111–128 (2014)
7. Essenwanger, O.: Curve Fitting. Wiley StatsRef: Statistics Reference Online (2014)
8. Bowden, R., Cox, S., Harvey, R., Lan, Y., Ong, E.J., Theobald, B.J.: Recent developments in automated lip-reading. In: Proc. SPIE 8901, Optics and Photonics for Counterterrorism, Crime Fighting and Defence IX; and Optical Materials and Biomaterials in Security and Defence Systems Technology X (2013)