A lip characteristic parameter-based Identity recognition algorithm

Yingjie Meng, Wei Chen*, LiXin Bai
School of Information Science & Engineering, Lanzhou University
Lanzhou, 730000, China
*Corresponding author Email: 243953106@qq.com

***Abstract*—most studies in lip-based speaker identity recognition fail to make full use of the changing characteristics of lips and have limitations of heavy information processing work and low rate of successful recognition. A new method of lip characteristics extraction and speaker identification based on the integrative analysis of both static and dynamic lip characteristics is proposed for this sake. The first step of this method is to extract the speaker's lip image set and mark critical points; then figure up the function curve of the speaker's mouth-corner points and underlip's lower midpoints, the function curve of the speaker's left mouth-corner's outer edge, the left highest points of upper lip's top edge and the midpoints of upper lip edge according to Lagrange Interpolation; then use the two kinds of coefficients integratively as movement characteristic parameters to analyze the similarity level between the unidentified speaker's lip characteristics extracted with that in the sample bank and the speaker can then identified. The results of simulation experiment prove that this method guarantees a high rate of successful recognition as well as lower complexity of computation and thus enjoys a good usability.**

Keywords- interpolation, characteristics extraction, lip motion

I. INTRODUCTION

Biological feature identification technology are those technologies that recognize personal identification through biological characteristics inherent in the human body (such as the retina, ear, lip etc.), which is approached by means of mainly computer-assisted acoustic, optical, biostatistics and biosensor technologies such as binding, use of and behavioral characteristics (such as gait, handwriting, voice etc.). it has now, out of its merits of stability, uniqueness and universality, become an important means of identification and is rapidly developed and widely used around the globe. The identity recognition based on lip movement as a biological characteristic among these enjoys a great potential since it is relatively simple in data collection and low in equipment cost. Lip feature information extraction is the most crucial step [1]. There are basically two ways of extraction, namely static approach and dynamic approach.

The static methods, as showed in bibliography [2, 3, 4], aims to split the lip motion video information into the image, and then the image size, shape, lip texture extraction as a lip feature information, then carries on the comprehensive utilization of the pixel information. It is easy to implement but limited in extracted information.

The dynamic approach, as showed in bibliography [5] strives for detection of lip motion parameter and the laws existed in it, and then analyze and process the data of lip motion characteristics. It is effective and can eliminate individual differences, but is sensitive to the rotation, scaling and light change. There is, however, a high feature vector redundancy and brings large amount of computation to later model training in spite of dimensionality reduction.

In consideration of their limitations, this paper will make integrative analysis of both static and dynamic features as speaker's lip feature information, and use Lagrange interpolation method to extract crucial lip characteristic parameters from key point in the lips, and analyze similarity with that in sample database for recognition of speaker's identity.

Based on existing studies [6], the present study set six sounds/vowels of {a; o; i; u; sh; z} as key factors, which can fully cover the speaker's lip feature information and reflect the speakers' individual differences to a large extent.

II. THE RECOGNITION MODEL STRUCTURE BASED ON THE INTERPOLATION OF LIP FEATURE PARAMETERS

Lip feature-based speaker's identity recognition generally consists of three phases: preprocessing, feature extraction and matching.

Preprocessing is mainly to select the key frames of speaker's video information, and then detect face image for located lip image sequence of the speaker. Such skills will not be explained in detail here since much mature achievements have been made in existing researches. The following part will be a discussion on the hypothesis that the speaker's sequential lip images of spelling the above six key factors has been obtained.

A. Lip feature information extraction

The process is as follows: mark on the key points on the lip image to get their coordinates first, and then, based on Lagrange interpolation, calculate the function curve coefficient along the key point on the lip image respectively, and two groups of curves are then combined to form the information of the speaker's lip motion features.

A lip feature information extraction model can be reached from the above description of the lip feature information extraction process.

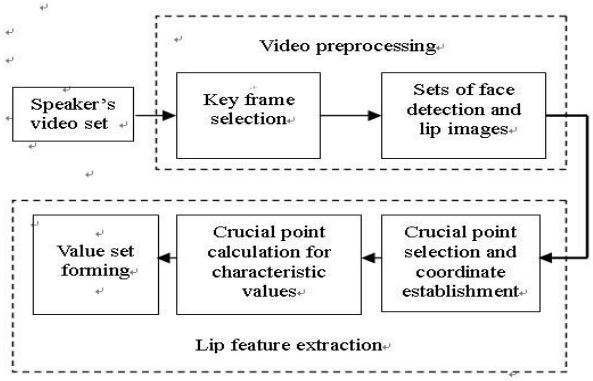

Fig.1 the process of lip feature extraction

B. *Lip feature-based identity matching and recognition model*

Lip feature based identity matching and recognition can be reached if a previously stored lip feature database is available. The process is as follows: to extract firstly the speaker's lip feature information ready for check with the method showed in figure 1; then get the candidate set by comparing the extracted lip feature parameters with that in the database according to the threshold parameter settings; the recognition results can be approached with a followed ordering of the obtained candidate set. This is the recognition model of figure 2 (also mentioned above). As for the unidentified speaker, we can enter their lip feature information into the database by machine learning as enrichment of its sample amount.

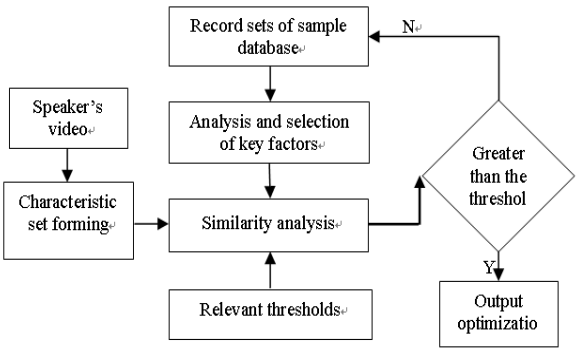

Fig.2 Identity recognition model

III. LIP FEATURE INFORMATION EXTRACTION

The main work of lip feature information extraction is to choose crucial points of the lip images and establish the coordinates according to speaker's lip image set obtained in the video preprocessing stage, and then make simulative function interpolation and we can get correlational coefficients, which means the speaker's lip motion characteristic information in key factors spelling. We can then get the speaker's lip motion parameter set. This is the whole process of lip feature information extraction.

The lip feature information extraction process can be designed as:

Step1: to obtain the speaker's lip image sets in the video preprocessing stage;

Step2: to select crucial lip points of the lip images and establish coordinates;

Step3: to make simulative function interpolation of crucial points in each image frame;

Step4: to get correlational coefficients as the speaker's lip motion characteristic information in key factors spelling;

Step5: to get the speaker's lip motion parameter set

A. *Crucial lip points selection and contents of characteristic information*

Since the two sides of human beings' lips are basically symmetrical, we can choose the two lip corners, the outside middle point of lower lip, the left highest point of the upper lip's two outside curves and their intersection point as five lip crucial points, as can be seen in figure 3.

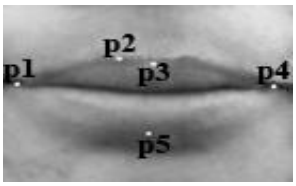

Fig.3 Example of lip crucial points selection

Since images obtained in the preprocessing stage is time-sequential and strong similarity exists between adjacent frames, we can locate the crucial point in the frame according to its position and pixel information in the previous one.

Motion feature extraction means to take the sequential images of key factors spelling from initial pronunciation to mouth close as research object and standardize the candidates coordinates of key points (as showed in figure 3), and obtain the respective curve coefficients of {P1,P2,P3} and {P1,P5,P4} through interpolation calculation. Two groups of curve coefficients can be taken as the speaker's lip motion characteristic information.

B. *Extraction algorithm of feature information*

Let abscissa value x[1.. c] be the interpolation nodes of crucial points, their function value be the ordinate value, and use key point tracking positioning algorithm of Position_ Key Point(V) in the existing researches[7], Motion feature

extraction algorithm G_Motion(V) can be designed as follows:

G_ Motion (V)

Input: Image sequence V for speaker's critical factor spelling

Output: movement characteristics for speaker's critical factor spelling

Begin

Step1: statistical record of sound numbers c;

Step2: use algorithm of Position Key Point (V) to mark crucial points of number u sound input image; //0<u<c

Step3: use Lagrange formula:

$$L(x)=\sum_{j=0}^{k} y_j \prod_{i=0,i\neq j}^{k} \frac{x-x_i}{x_j-x_i}$$ to start Lagrange interpolation of the crucial points of sound u's number w image and get interpolation coefficients (a0…a5);

Step4: get the average $\bar{a}_j=\sum_{i=1}^{14} a_j/14$ (j=0…5) of interpolation coefficients of all sound number u's crucial frame images as the sound's characteristic parameters H;

Step5: continue step 2 if u<c and stop otherwise;

Step6: go back to the speaker's pronunciation characteristic parameters;

End

IV. THE DESIGN OF LIP CHARACTERISTIC PARAMETER-BASED IDENTITY RECOGNITION ALGORITHM

The main task of lip characteristic parameter-based Identity recognition is to: first use the method mentioned above to obtain the speaker's lip feature parameters, and then analyze their similarity with that in the database, followed by screening with threshold to get candidate set and optimize for identification. Input the speaker's lip characteristics into training model for further recognition if the candidate is empty. The process of speaker identification is as follows:

Step1: form the lip characteristic parameters of the speaker to be identified;

Step2: extract one lip characteristic parameter from the sample database;

Step3: form key factor set by analyze lip characteristic parameters;

Step4: comparing with the key factor set of the speaker's characteristic parameters;

Step5: calculate the result and put into candidates if it is greater than the threshold;

Step6: turn to step2;

Step7: optimize and get recognition results;

Step8: train and input the speaker's lip characteristics into the sample database if the result is less than the threshold or empty.

A. Similarity analysis

Similarity analysis is used to calculate the speaker's lip characteristic parameters with that in the sample database and find out to what extent they are similar, and use this coefficient as standard for judgment. Here we use correlation coefficient of two vectors for analysis.

Take the characteristic parameters as one-dimensional vector and com as similarity level of two vectors, vector $s(s_1,s_2,\ldots,s_n)$ as the information of speaker's lip characteristic parameter in the database and vector $t(t_1,t_2,,\ldots,t_n)$ as that to be identified, the specific calculation formula is:

$$\text{Com}(s,t)=r(s,t)=\frac{\sum_{j=1}^{n}(s[i]*t[i])s,t)}{\sqrt{\sum_{j=1}^{n}s[i]^2}*\sqrt{\sum_{j=1}^{n}t[j]^2}} \quad (1)$$

The similarity analyzing process can be described as:

Proc_ similarity (s, t, VAR com)

Input: two vectors of $s(s_1,s_2,\ldots,s_n)$ and $t(t_1,t_2,,\ldots,t_n)$

Output: similarity (com)

Begin

Step1: record vector length (n);

Step2: Initialize variables in the calculation and let Total1, Total2, Total3 are zero;

Step3: let j=1 and make partial calculation of formula (1);

Total1←Total1+ s[j]*t[j]; //calculate ∑s[j]*t[j];

Total2←Total2+s[j]*t[j]; // calculate $\sum s[j]^2$;

Total3←Total3+ t[j]*t[j]; // calculate $t[j]^2$;

Step4: let j=j+1, continue step 2 if j<n+1 and stop if not;

Step5: make similarity analysis according to formula (1);

Step6: go back to similarity com

End

B. Lip characteristic parameter-based Identity recognition

Let H' be the lip feature parameters extracted from video information of speakers to be identified, then we can select and input those speech, according to the similarity between H' and previously stored sample parameters segments D[1..m], into candidate set V if they are higher in similarity than the established threshold.

Let V1' be critical factors spelling image sequence of the speaker to be identified, D[1..M] be those in the sample database and T be the similarity threshold, the interpolation parameter-based identifying process can be described as follows:

Proc_detect (V1' [1...N, 1...n], D [1...m, 1...n], T)

Input: The total frame number of key factors to be detected (N), sequence V1' [1...N, 1...N], parameter set in

the collected sample database (D [1...M]) and threshold value (T)

Output: speaker's identity

Begin

Step1: use algorithm G Motion (V1') to extract characteristic parameter H';

Step2: integrate the characteristic parameters of the multiple pitch into one-dimensional vector a;

Step3: let i=1,j=1,b=D (i, 1...n) and start similarity analysis;

Step4: use algorithm similarity (a ,b ,com) to calculate the similarity value between a and b;

Step5: input the speaker's identification and similarity into candidate set V, j=j+1, i=i+1 if com>threshold T;

Step6: continue step 4 if i<m+1 and stop otherwise;

Step7: stop if j=1 which means output recognition fails and go on to the next step if not;

Step8: put candidate set V into descending order according to their similarity values;

Step9: output the first three speaker's identity information;

End

V. SIMULATIVE VERIFICATION EXPERIMENT AND ANALYSIS

Simulation experiment is used here to verify validity and usability of the extracted lip characteristic parameters. Experiment equipment: common PC machine, MS Windows XP, MATLAB 7, MS Visual C++ 6.0.

The sample data used in the experiment are from bibliography [8] and consists of 20 different speaker's six key factor spelling video data, 30 frames per second, 24 bit color and resolution 640*480. Sequential images from initial pronunciation to mouth close are extracted from the video data. Preprocessed lip images are 255 level grayscale and resolution 128*96, part of which are showed in figure 4.

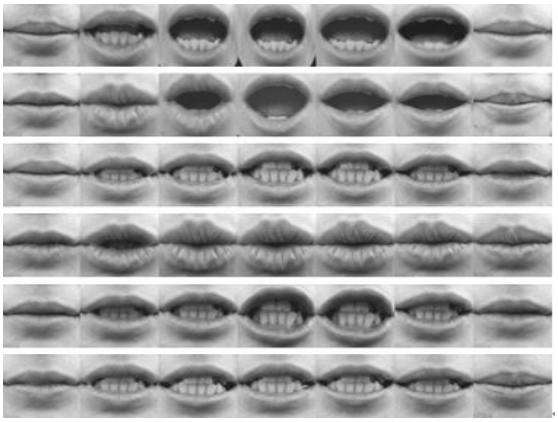

Fig.4 Examples of lip images in the sample database

A. *Similarity comparison with key factors*

We can compare the lip's pronunciation images with crucial point interpolation curve and then analyze the validity and usability of extracted characteristic parameters. Followed is an example of comparison.

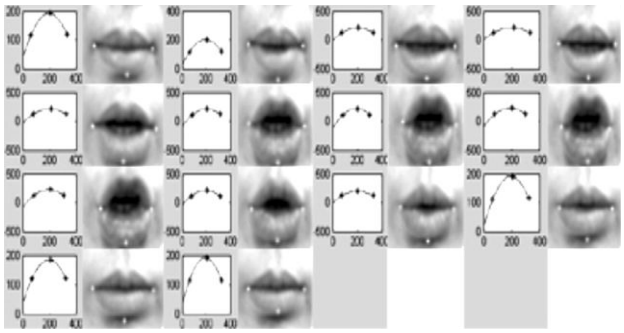

Fig.5 Contrast of lip images their curves in spelling o

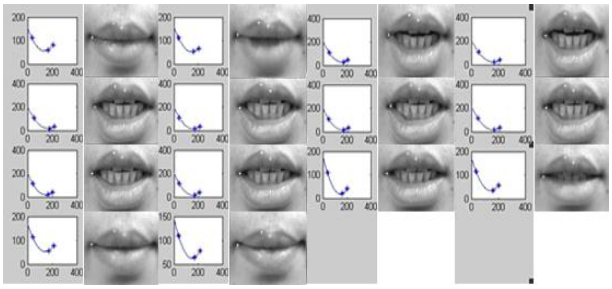

Fig.6 Contrast of lip images their curve in spelling sh

We can see from the above two figures that the crucial points interpolation curve are basically mirror symmetric with speaker's actual pronunciation, which is a proof that the present lip characteristic information extraction has a relatively good availability.

Followed is a table of one speaker's lip feature information parameters:

TABLE 1 SPEAKER'S LIP CHARACTERISTIC INFORMATION PARAMETERS

	Curve parameters of the upper lip			Curve parameters of the lower lip		
	A0	A1	A2	B0	B1	B2
a	207.038	-2.07847	0.00614882	14.3722	1.89415	-0.0045931
o	147.47	-1.34978	0.00411004	40.1778	1.71721	-0.0040864
i	272.879	-2.75051	0.00840238	-25.4541	2.35897	-0.005987
u	187.082	-2.01812	0.00650604	36.3382	1.88397	-0.0045995
sh	206.961	-2.35252	0.00791515	50.4919	1.76451	-0.004432
z	142.79	-1.38505	0.00451024	30.3359	1.69775	-0.004077

A0, A1 and A2 are coefficients generated from interpolation of crucial points of the upper outside lip while B0, B1 and B2 are that of the lower outside lip, with a, o, i, u, sh, z as key factors.

B. *The rate of successful recognition*

The rate of successful recognition is the correctness percentage of identification to all speakers in detection, which is of course affected by many factors, mainly the model structure, recognition algorithm, and the speech database etc.

Using 20 speakers' lip voice and video as samples, the present paper compares first the crucial points number and similarity level between the lip characteristic information of speakers to be identified with that in the sample database first and then the present algorithm with that used in bibliography [9] and [10] (the rate of successful recognition are from bibliography [8]). Table 2 is the results:

TABLE 2 RESULT COMPARISON OF SPEAKER IDENTIFICATION

	Number of crucial points	Rate of successful identification
Bibliography [9]	11	78.79%
Bibliography [10]	9	75.76%
The present algorithm	5	75.16%

As can be easily seen from the table, the present method enjoys a relatively higher paper has a higher rate of successful recognition, which is a sound proof that crucial point selection and interpolation are usable. Fewer crucial points' selection than both bibliography [9] and [10] also means smaller time and space consumption in calculation.

VI. SUMMARY

This paper presents a new lip feature information extraction method and integrates both static and dynamic data as the speaker's lip feature information, making it more comprehensive than using either of them alone. This ensures a higher rate of successful speaker identification. As for the complexity of calculation, five crucial points are selected to mark lip features here and their extracted parameters are compared as well as interpolated, followed by similarity analysis with that in the sample database. Application of Lagrange interpolation lowers effectively the complexity level if contrast with the method used in bibliography [9] and [10]. Simulative experiment shows a relatively good validity and usability of the present algorithm, which has also certain reference significance for lip reading technology and animation modeling other than speaker identity recognition. Further efforts can be made in research of its practical application.

REFERENCES

[1] Çeting ül. H. Ertan, Yücel.Yemez1, Erzin. Engin, Tekalp.A. Murat. Discriminative analysis of lip motion features for speaker identification and speech-reading. IEEE Transactions on Image Processing, v 15, n 10, p 2879-2891, October 2006.

[2] J. Luettin, N. A. Thacker. Speechreading Using Probabilistic Models. Computer Vision and Image Understanding, v 65,n 2,p 163-178, February 1997

[3] G. I. Chiou, J. N. Hwang. Lipreading by Using Snakes. Principal Component Analysis and Hidden Markov Models to Recognize Color Motion Video. IEEE Trans. on Image Processing, v 6,n 8,p 1192-1195, 1997

[4] T.F.Cootes, C.J.Taylor. A Mixture Model for Representing Shape Variation. Image and Vision Computing, v 17,n 8,p 567-573, June 1999..

[5] Li Meng,Cheung Yiu-Ming.A Novel Motion Based Lip Feature Extraction for Lip-reading. Proceedings - 2008 International Conference on Computational Intelligence and Security, CIS 2008, v 1, p 361-365, 2008

[6] LI,Jing;Zheng,Fang;Zhang,Jiyong;Wu,Wenhu. Context dependent initial final acoustic modeling for continuous Chinese speech recognition . Qinghua Daxue Xuebao/Journal of Tsinghua University, v 44,n 1,p 61-64,January 2004..

[7] Meng,Yingjie;Bai,LiXin;Liu,WenJun;Liu,MingWen. Extraction algorithm of lip characteristic parameters based on interpolation . Applied Mechanics and Materials (Advances in Mechatronics, Robotics and Automation II), v 536-537,p 235-240,2014

[8] LiXin Bai, Speaker-identification Algorithm Research based on Lip-features[Master's thesis].Lanzhou: Lanzhou University,2014.

[9] Meng Yingjie, Li Zhaoxia,Hu Yingjie, et al. Speaker identification based on feature mouth shapes.Journal of Information and Computational Science,v 6,n 3,p 1209-1216,June 2009..

[10] Meng Yingjie, Hu Yingjie, Zhang Haiyan, et al. Feature mouth shapes extraction based on contour of internal lips.2010 6th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2010. September 23, 2010 - September 25, 2010