Analysis and Determination of Inner Lip Texture Descriptors for Visual Speech Representation

Xibin Jia
College of Computer Science, Beijing University of Technology, Beijing, China
Email: jiaxibin@bjut.edu.cn

Hua Du and Yanfang Han
College of Computer Science, Beijing University of Technology, Beijing, China
Email: dhmj2012@emails.bjut.edu.cn, hyf1107@126.com

David M W Powers
College of Computer Science, Beijing University of Technology, Beijing, China
School of Computer Science, Engineering & Mathematics, Flinders University, Adelaide, Australia
Email: david.powers@flinders.edu.au

***Abstract*—The problem of visual speech representation for bimodal based speech recognition includes particular challenges in the modeling of the inner lip texture reflecting different pronunciations, such as the appearance of teeth and tongue. This paper proposes and analyzes several possible statistical inner lip texture descriptors to determine an effective and discriminant feature. Simply using grayscale without full specification of the underlying colour model tends to loss some significant discriminative information. Therefore thorough exploration on the color space components selection in computing the local inner lip texture is thus a primary goal of the present research. The L channel of Lab color space is finally determined as the basis for the development of the inner lip texture model. Through feature level fusion, the final classification of visual speech is performed based on the proposed inner lip texture descriptor and standard geometric features. Together with audio speech, this paper furthers the development of the CHMM based bimodal Chinese character pronunciation recognition system. The experimental results show that the local inner texture descriptors, such as the color moment with geometric feature, outperform the holistic inner texture descriptors, such as the statistical histogram, in representing visual speech with the close discriminability but low dimensionality.**

***Index Terms*—inner lip texture descriptor, local feature, feature fusion, visual speech representation**

I. Introduction

Speech recognition is one of the fundamental technologies in natural human computer interaction. Traditional speech recognition technologies have achieved higher recognition rate of words or phrases for a noise environment. However, for use in a realistic environment, the recognition performance is compromised because of factors such as the background noise and the complexity of the auditory and reverberatory environment[1,2]. Indeed psychological research results suggest that only 11% percent of information comes from audio speech and 83% from the visual modality for human speech understanding[3]. Thus, bimodal or multimodal speech recognition is critical for effective improvement of the rate and robustness of speech recognition, as it is necessary to use both the complementary information (one modality strongly better) and the redundant information (balanced strength across modalities) in order to exploit their different artefact and noise properties – what is redundant under perfect conditions can provide critical information under noisy, degraded or impoverished conditions.

Traditional speech recognition focuses exclusively on auditory information[4], while this paper concentrates on improving the ability to extract visual speech information. Currently there are two major categories to visual feature for speech.

The first class of algorithm focuses on geometric-based feature such as lip contour. This category of techniques employ algorithms such as ASM (Active Shape Model) [5], snake algorithms [6] and AAM (Active Appearance Model) [7] to track interested points or contours to compute the geometric feature such as lip width and lip height (of either or both the outer and inner lip). Zhang [8] shows that inner contour information has good performance in representing visual speech by improving the recognition rate of 11%. Zhi [9,10] makes comprehensive comparison among geometric features such as pixel counts for outer lip width (X), height (Y), area of outer lip surround region, along with the angle at the lip corners (θ), the outer aspect ratio Y/X, and first-order derivatives or differences of the above features. The experiment results prove that combination feature such as X-Y-dθ/dt gives the best recognition results.

The second kind of approach is based on the texture of lip region images as determined using holistic features

doi:10.4304/jcp.9.7.1628-1638

such as PCA (principle component analysis), DCT (discrete cosine transformation) and DWT (discrete wavelet transformation) [11,12]. But these approaches have higher dimensionality and are not specialized for visual speech representation or the exploitation of redundancy in revealing the visual speech characteristics.

Chan [13] presents a hybrid technique that makes use of mouth height and width plus a normalized intensity profile as a surrogate for detailed tooth and tongue tracking, showing immense sensitivity to normalization by both average intensity and tracking contour bounding box (in a small single-speaker isolated digit recognition task). Mersereau [8] also discusses the importance of state of teeth and tongue in describing the speech, using two additional components to express the teeth and tongue states with 0 representing invisible and 1 visible teeth or tongue, with the recognition rate increase 5%.

Eric Benhaim etc.[14] proposed a novel visual speech representation based on histogram-based descriptors of local interested region by computing histograms of oriented gradient (HOG) and histograms of oriented optical flow (HOF). Multiple kernel learning is employed in selection and combination of a discriminant (appearance and motion) structured features of different orders. Topkaya [15] etc. uses multiple classifiers to obtain multiple visual tandem features. In these approaches, discriminant classifiers are one of key techniques to improve the understanding of visual speech.

Geometric features have the advantage of low dimension and robustness to both face gesture and illumination. But, it cannot reveal speaking states such as the visibility of tongue or teeth, which is important in describing speech. Holistic features have rich and complex information but it is not optimized for speech description, whilst the local texture can correlate with aspects of the target states but missing the outline.

In view of the limitations of each technique alone, this paper adopts a joint feature vector with geometric and local texture information combined. We specifically explore the local texture representation of the inner lip together with general geometric features in describing the opening width and height. Several possible inner texture features are proposed or used and comprehensive comparing experiments are done to determining the better form. We thus used several different features in our CHMM (couple hidden Markov Model) bimodal speech recognition system to analyze the effectiveness of joint features.

The remaining parts of the paper are as follows. Section II introduces our computational model of joint visual speech features. Section III provides details of comprehensive comparison experiments for choosing an effective inner lip texture representation and determining the adopted color space components. Section IV introduces our CHMM-based bimodal Chinese character pronunciation recognition system and explores the performance of possible visual speech features. Conclusions and Future Work are discussed in the final section.

II. Joint Visual Speech Feature Based on Geometric and Inner Lip Texture

A. Our Geometric Visual Speech Feature

Form of our geometric visual speech feature: As the phonetic states are mainly reflected in the mouth area with the change lip shapes, we adopt the most frequently used parameters to construct visual speech geometric feature that reflect the extent of opening and rounding of lips, including the outer and inner width, height and angle formed at the mouth corners. For simplicity we removed the repetitive symmetric components, although these may be of use in specific, noisy or abnormal cases. We finally settled on a 7-dimension feature vector representing the geometric feature $\boldsymbol{F_G}$ in terms of the tuple (θ, w_2, w_1, h_4, h_3, h_2, h_1) shown in Figure 1, where θ is the angle of the outer lip corner, w_2 is the width of the outer lip, w_1 is the inner width, h_4 is the outer height, and h_3, h_2 *and* h_1 represent the inner height at three increasingly lateral locations. Note that the precision of the location of outer lip points is relatively higher while that of inner lip is lower, influencing by the subtlety of the change of inner lip texture as faced by the AAM algorithm. Thus we adopt the three inner height parameters, not only to provide further shape information but to mitigate the error influence, and increase the stableness of inner lip opening extent description with complementary and redundant information while retain only one component for outer lip height.

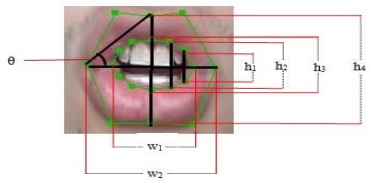

Figure 1. Illustration of Geometric Feature

AAM for computing the geometric feature: The paper uses the AAM algorithm in AAMlib to extract the feature points in the face images. Using the extracted prenasale as the reference point, we segment the lip region shown in Figure 2.

Figure 2. Illustration of Lip Region Image Partition

Taking the fact into account that the changes of the inner texture brings the lower accuracy of point location, the paper proposes to training multiple AAM templates instead of a unique AAM template based on the lip shape types. Analyzing the mismatching cases with the unique AAM template (see results shown in Figure 3), it can be seen that the mismatching is mainly caused by the following two situations, shown in Figure 4. The first is the mouth closing state and the second is the round mouth shape with inner corner points are away from the outer ones. Therefore we propose training three AAM

templates respectively with separating the round and close speaking mouth shapes from the others where we designate as ① closed mouth AAM template, for the phonemes such as "b,p,m,f" in Chinese, ②rounded open mouth "O" type AAM template for phonemes such as "o,ou,ong" in Chinese, and ③ general AAM for the remaining lip shapes. The representative lip shapes are shown in Figure 4.

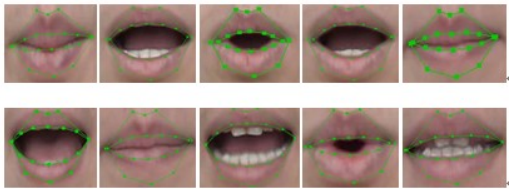

Figure 3. Point Location Results based on single AAM Template

Figure 4. Representative lip shapes for three AAM templates

Figure 5 shows point location results using this cluster various AAM templates. We can see that the location of inner corner points improved when the outer and inner ones are not overlapped due to the use of separate AAM template rather than the single template. The illustration also makes obvious the accurate inner lip point location in closing mouth states.

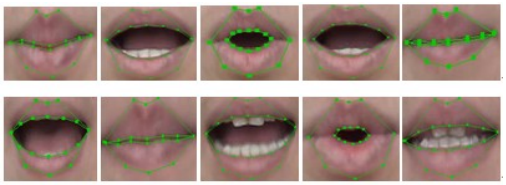

Figure 5. Point Location Result based on Cluster-Various AAM Templates

B. Texture Feature of Inner Lip

Partition of inner mouth area image: The inner contours are obtained by simply connected the extracted inner points as boundary. The pixels outside the boundary are removed by assigning a constant pixel value, here we use 255. Images of inner lip area are generated as shown in Figure 6.

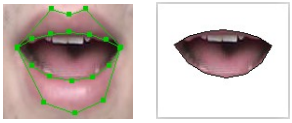

Figure 6. Illustration of inner lip area generation

Multi-feasible local inner texture features: Referring to some general texture descriptors and combing our understanding of pronouncing inner lip texture states, we explore the following inner texture features.

①Histogram of inner lip region. This feature describes the distribution of gray/color in the inner lip region. To eliminate differences such as mouth size of individuals, we use the normalized histogram feature representing in (H_0, H_1, ... H_{255}), where each component is obtained by dividing the total pixel number in inner lip area as in formula (1). As we count frequency of every gray/color scale, the dimension of this gray/color histogram feature is 255.

$$Hi = \frac{N_i}{SUM} \tag{1}$$

where $SUM = \sum_{i=0}^{255} N_i$, $0 \le i \le 255$, and N_i is the number of pixels with pixel value i, SUM is total number of pixels inside the inner lips.

As this 256 grayscale histogram is a feature with higher dimensionality, we seek to establish a compact feature of low dimension histogram by combing several continuous grayscale as one scale. In the paper, we use the 32 grayscale/color histogram with 8 continuous pixel values as one scale.

②Block Histogram. This feature is proposed in order to represent visibility of upper and lower teeth separately. Actually the different visibility of upper and lower teeth is relative to the different pronunciation, but the holistic histogram can't discriminate this phenomena. So we separate the inner lip region into upper and lower parts (shown in Figure 7)to compute the histogram respectively and concatenate them into 512 or 62 dimension histogram or compact histogram feature vectors with the above form we call it block histogram feature.

Figure 7. Partition Blocks of Inner Lip Region

③Teeth appearance proportion: teeth in the inner lip region are salient comparing with the rest area and robust to lighting relatively. The visibility extent of teeth also reflects the different pronouncing states. So the paper proposes a kind of inner lip pronouncing texture descriptor with the teeth appearing proportion. Actually analyzing the histogram of inner lip texture images with and without teeth in Figure 8, we could find that grayscale of teeth locates in range [200,400] while non-teeth's grayscale locates in range around [120,200]. Thresholding approach is employed in the paper to extract the teeth area. Then the teeth area proportion and also the upper and lower teeth area proportion are counted to construct the three dimension local inner lip texture descriptor.

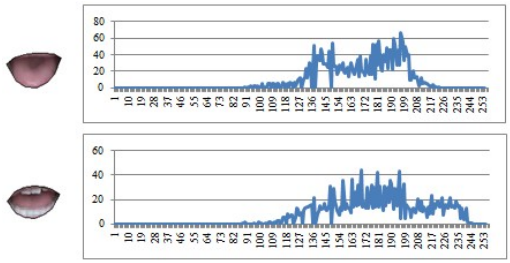

Figure 8. Comparison of inner lip texture distribution of Pronouncing Speech with and without teeth appearance

④ Blockaverage. This feature is established by partitioning the inner lip region averagely into 8 sub-region and counting the average grayscale/color of pixels in each sub-region representing in b_i for ith block, then concatenating into 8 dimension vector (b_1, b_2, ... b_8) as the inner lip texture descriptor. Formula (2) illustrates the counting of each component b_i, where N_i is total pixel number in ith sub-region, X_{ij} is grayscale/color value of jth pixel in ith sub-region.

$$b_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij} \quad \text{where} \quad 1 \le i \le 8 \tag{2}$$

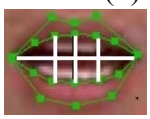

Figure 9. Ilustration of Partition of Inner Lip Area

⑤ Grayscale/Color moment. Color moment is one of common texture descriptors and good at reveal the local texture changes. Considering the texture change involving with the pronunciation mainly happens along the vertical direction, we establish a kind of inner lip texture feature by counting the first color moment in vertical direction. To mitigate the noise influence, we count the color moment of three local regions shown in Figure 10. And establish the three dimension feature (c_1, c_2, c_3) with color moment. The computing formula of each component is shown in formula (3). Where x_{ij} is value of grayscale/color of jth pixel from upper to down in the ith vertical, S_{ij} is the index of this pixel counting down from the top along the ith vertical, N_i is total amount of pixels in ith vertical that lie inside the inner lip.

$$c_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij} \cdot S_{ij} \quad \text{where} \quad 1 \le i \le 3 \tag{3}$$

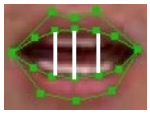

Figure 10. Ilustration of Computing Region of Color Moment

⑥DCT(Discrete Cosine Transformation)：One of the most frequently used holistic transformation feature is the DCT, which we therefore use as one of local features in our paper to explore its capability of representing phonetic information. It is computed as formula (4) for a $M \times N$ image with pixel value fx,y in each coordinator (x,y).

$$C(u,v) = a(u)a(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f_{x,y} \cos\frac{(2x+1)u\Pi}{2M} \cos\frac{(2y+1)v\Pi}{2N} \tag{4}$$

where $a(u)$ is:

$$a(u) = \begin{cases} \sqrt{\frac{1}{N}} & if \ u = 0 \\ \sqrt{\frac{2}{N}} & if \ u = 1,2,3 \ldots N-1 \end{cases}$$

The energy of DCT mainly concentrates in top-left of the transform plot shown in Figure 11, so we use just the first 32 dimension of the DCT of inner lip area as our DCT features.

Figure 11. Example of DCT Transformation Result

Color component for inner lip texture: Based on the definition of the above feasible features, it is appropriate to explore and choose an efficient color channel in the feature computation, to get a better discriminative inner texture descriptor for speaking states.

From the perspective that inner texture information is easily influenced by the illumination, the paper prefers to choose a color space component that is robust to illumination changes. Taking the lip images in Figure 12 as an example, we compare our images in each chromatic channel in some of the most frequent used color spaces, viz. HSV, Lab and RGB, shown in Figure 12. The discriminability of inner lip texture varies among those color spaces. For example, the S channel image in HSV color space, which represents the strength or saturation of the colour, has given more obvious contrast between the region of white teeth and pigmented skin or inner mouth tissue. Conversely, the H namely hue component is of little value as the skin and tissue is basically a filtered red or blood color [16] and useful for finding faces but not for finding features within faces. It could be inferred that selecting the S component to building the inner lip texture may be helpful to reflect the difference of speaking mouth inner lip texture, while the H component would not.

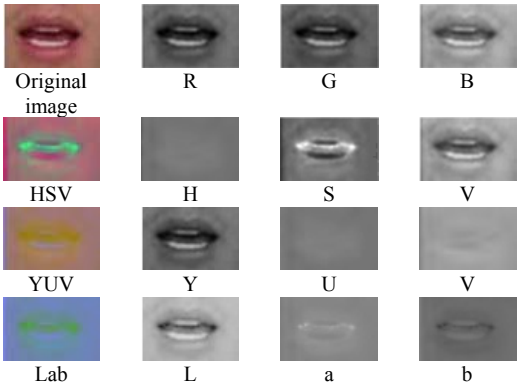

Figure 12. Example of Color Channel Image in differnet Color Spaces

For each color system, it uses the different principles to realize the description of the color space. RGB uses additive color mixing, describing what kind of light needs to be emitted to produce a given color, and is inspired by human vision capabilities and defined based on the characteristics of camera and display technology. RGB stores individual values for red (R), green (G) and blue (B) and is the most common form for describing colors for display. HSV (hue, saturation, value), is employed to describe the color of image in terms of hue and saturation rather than in terms of additive or subtractive color components, and is designed for ease of selecting colors by humans in color selector interfaces. HSV is a simple transformation of the RGB color space, with the R, G and B corresponding to hue angles of 60, 120 240 degrees (mod 360) although this is often renormalized to fit into a byte (mod 255), while the grayscale-like value V represents the strength of the strongest of the RGB components, and the saturation the maximal difference between the components (and hence difference from grayscale defined as all RGB components having equal value). The YUV model also called YCbCr in its digital version defines a color space in terms of one luma (Y, namely brightness) and two chrominance (UV) components, and is based on the opponent colour model of human vision, with a grayscale model weighted to reflect all RGB components, not just the most dominant, being designed to allow colour TV signals to piggy back just two low bandwidth colour difference signals onto a backward compatible mono-chrome signal, and this remains the underlying model for modern compression techniques such as JPEG. The Lab color space (and its more sophisticated derivatives) is a similar color-opponent space that is designed to more accurately reflect human perception under the full range of lighting conditions, as well as accurate mapping to and from particular devices, with dimension L for lightness and a and b for the color-opponent dimensions. One of the most important attributes of the Lab-model is device independence. This means that the colors are defined independently of the apparatus they are created with or the device they are displayed on. At the same time, unlike the RGB color models, as Lab color is designed to approximate human vision, it aspires to perceptual uniformity, and its L component closely matches human perception of lightness. It can thus be used to make accurate color balance corrections by modifying output curves in components a and b, or to adjust the lightness contrast using the L component. Thus in principle, the Lab color space is preferred to for inner lip texture, to enhance the discrimination of visual speech, which is clearly defined by human perception characteristics.

This gives us broad idea of the importance of selecting color components appropriately, and we explore this empirically in Section III.

C. Joint Visual Speech Feature Based on Geometric and Inner Lip Texture

This paper adopts a feature-level fusion approach to combine the geometric feature and inner lip texture feature together to generate the joint visual speech feature. Supposing that the geometric feature G is marked as $(g_1,g_2.....g_m)$ and the texture feature $\boldsymbol{T}$ as $(t_1,t_2.....t_n)$, the joint feature $\boldsymbol{F}$ is the concatenated vector of $\boldsymbol{G}$ and $\boldsymbol{T}$ as follows: $\boldsymbol{F}=(\boldsymbol{G},\boldsymbol{T})=(g_1,g_2.....g_m,t_1,t_2.....t_n)$.

To tradeoff the metric difference and balance the contributing weight, normalization is done as a preprocessing step. The common normalization techniques include Z-Score, Media-MAD, Min-Max etc. The simple Min-Max normalization technique was tested outperforming the others [17], and is adopted for this research on this basis. The normalization procedure for the joint feature $\boldsymbol{F}$ is performed as follows, for the visual speech image sample set $\{\boldsymbol{X}_i|i=1,2.....N\}$.} Counting up the maximum and the minimum of each component of the joint visual speech feature $\boldsymbol{F}_i=(\boldsymbol{G}_i, \boldsymbol{T}_i)$ in the sample set, denoting as G_{max}, G_{min}, T_{max} and T_{min}, the normalized feature of sample $\boldsymbol{X}_i$ is $\boldsymbol{F}_i^{'}$ in formula (5)

$$\boldsymbol{F}_i^{'}=\boldsymbol{G}_i^{'},\boldsymbol{T}_i^{'}=(g_{1,i}^{'},g_{2,i}^{'}...g_{m,i}^{'},t_{1,i}^{'},t_{2,i}^{'}...t_{n,i}^{'}) \quad (5)$$

where: $g_{m,i}^{'}=\dfrac{g_{m,i}-G_{min}}{G_{max}-G_{min}}, t_{n,i}^{'}=\dfrac{t_{n,i}-T_{min}}{T_{max}-T_{min}}$

III. ATTRIBUTION ANALYSIS EXPERIMENT OF VISUAL FEATURES

A. Data Collection

For this research we collected a dataset for one subject for analyzing the performance of the above features of the inner lip texture defined in section 1. The corpus is selected from some reading materials based on the phoneme balance rules to cover most Chinese phonemes. Part of corpus we used is as Table I. There are 100 sentences in our corpus. The data is collected using Sony DMC_LX3 camera with resolution of 720x570 and frame 25fps. We capture the front face data of a Mandarin speaker while reading the materials. The collected video was manually segmented into clips of single Chinese phonemes (Consonant and vowel) with Movie maker. The corresponding frame sequence images are stored as the sample data.

TABLE I
PART OF CORPUS SET

我是北京工业大学学生(I am a student of Beijing University of technology)
wo shi bei jing gong ye da xue xue sheng.(In Pin Yin)
走到半路天色暗了下来(It gets dark on the way)
zou dao ban lu tian se an le xi alai(In Pin Yin)
这件任务使他有点儿手足无措(This task makes him bewildered)
zhe jian ren wu shi ta you dian shou zu wu cuo. (In Pin Yin)
琳决定宴会一开始就分发礼物(Lynn decided to distribute gifts at the beginning of banquet.
lin jue ding yan hui yi kai shi jiu fen fa li wu. (In Pin Yin)
我和我女儿路过这儿上摩天岭(My daughter and I will go to Motian Ling Mountain through here)
wo he wo nv er lu guo zhe shang mo tian ling(In Pin Yin)

According to lip shapes of Chinese viseme pronunciation and referring to the definition of Chinese viseme cluster proposed by Zhao and Tang [18], we adopt the 9 Chinese viseme cluster shown in Table II. The number is viseme cluster number and relative Chinese phonemes. This is employed into the labeling of the above data collection.

TABLE II
DEFINITION OF CHINESE VISEME CLUSTER

Cluster No of viseme	Relative phonemes	Cluster No of viseme	Relative phonemes
1	b,p,m,f	6	o,ou,ong
2	d,t,n,l,g,k,h	7	e,er,ei,en,eng
3	j,q,x	8	u,v,un
4	z,c,s,zh,ch,sh,r	9	i,in,ing
5	a,ai,ao,ang,an		

We didn't use all the images in the Chinese phoneme pronouncing sequences, but selected the keyframes with the salient pronouncing states and abandoned the transitional images. For the single vowels viz. 'a' and simple compound finals viz. as 'ang', we selected multiple salient images representing the varied lip shapes and label them with relative cluster labels. For the complex compound finals such as 'ia,iao,ian,iang,ie,ua,uai,ui,ue,iou,iong,uo' with multiple pronouncing lip shapes, we chopped them into several here two segments and do the salient frame selection and labeling then. For example, 'ia' is separated into 'i' and 'a' clips and labelled with cluster No. 9 and 5 separately. We finally have 450 labeled images in our sample dataset.

B. The Experiment Results and Analysis

On the labeled dataset, we employ the optimized support vector model, SMO, as well as BayesNet as classifiers respectively to realize the Chinese viseme recognition and do the following analysis: 1) Analyzing the performance of color components in computing the inner lip texture. 2) Analyzing the discriminability of the above inner lip texture. 3) Analyzing the performance of the joint geometric and each inner lip texture. During the experiments, we use 10-fold cross-validation (CV) to allow reliable assessment without the need for expensive hold out data. The experiments are done as follows.

We first analyze the performance of color components in computing the inner lip texture. Here we use the four color systems: RGB, HSV, Lab and YCrCb. The inner lip texture features described in section I (Colormoment, Blockaverage etc. which are listed in Table 3) are computed respectively based on each color channel to represent the speaking mouth images. Training and testing the classification results in the collected dataset with identifying the 9 viseme clusters, we count up the

TABLE III
SPEAKING LIP SHAPE RECOGNITION RESULTS BASED ON EACH COLOR SPACE COMPONENT WITH DIFFERENT INNER LIP TEXTURE FEATURES.

Color channel	Colormoment		Blockaverage		DCT	
	SMO	BayesN	SMO	BayesN	SMO	BayesN
H	15.1	22	43.4	49.1	49.5	45.6
S	34.8	36	54	56	62	57.6
V	41.5	33.4	41.7	44.8	60.2	58.3
L	39.2	36.9	52.1	54.8	63.7	59
a	22	33.6	49.2	48.3	50.7	49.8
b	26.9	29.2	47.5	46.2	55.8	54.2
Y	45.1	38.2	48.7	49.9	62.5	58.3
Cr	14.4	23.2	47.5	30	40.7	37.3
Cb	20.4	30	41.2	45.7	44	40.3
R	45.2	37.8	48.7	51.5	63.1	57.6
G	42.9	38.5	48.4	51.7	62	58.1
B	41.5	33.4	48.7	51.2	60.6	58.3
Gray	40.7	36.5	48.3	49.9	59.5	57.4
Average1 (%)	33.1	33.0	47.6	48.4	56.5	53.2
Average2 (%)	33.0		48.0		54.9	

Color channel	Histogram		Blockhistogram		Average (%)
	SMO	BayesN	SMO	BayesN	
`H	25.6	29.1	32.1	28.5	34
S	48.8	53.7	57.4	65	52.5
V	58.5	60.9	61.8	67.3	52.8
L	59.2	67.5	62.3	71.8	**56.7**
a	50.5	51.4	54.9	53.2	46.4
b	57.4	57.8	61.1	61.3	49.8
Y	56.9	64.3	62.9	68	55.5
Cr	42.1	41.4	47.2	48.1	37.2
Cb	49.5	52.3	53.7	54.1	43.1
R	56.7	60.4	57.8	67.3	54.6
G	52.3	58.6	57.4	67.3	53.7
B	58.3	60.8	61.8	67.3	54.2
Gray	51.2	57.3	56.7	65.2	52.3
Average1 (%)	51.3	55.0	55.9	60.3	
Average2 (%)	53.2		**58.1**		

accuracy rates that the tested visual speech images are correctly classified into the right viseme clusters. Then we discuss the discriminant performance of the color space component and determine the one as the basis for computing inner lip texture feature. Note that teeth proportion feature is based on thresholding approach, the boundary of teeth and tongue region is hard to be found and determine the threshold, so we don't list teeth proportion feature in these color channels with incomplete data.

The experiments results are shown in Table III (with best results indicated in bold). We can find that the average recognition accuracy rates of each color channel show that L component in Lab color space has the highest rate 56.7% in different inner texture feature comparing with the other channels. This accords with our understanding discussed above that definition of L is matched to human perceptual characteristics. Therefore, we select the L channel as the basis to compute the inner lip texture. From the perspective of the texture feature, we could find that the block histogram has the highest average recognition rate, it proves that upper and lower texture contribute different in telling the inner texture of visual speech. Another two holistic features: DCT and Histogram also shows better discriminability comparing to the local ones when using the inner texture feature only. But the effectiveness still needs to be considered together with geometric features, for local texture features lose too much information about lip shapes.

We also test the effectiveness of our proposed compact histogram representations histogram-32 and blockhistogram-64 in Table IV compared with the uncompact ones: Histogram-256 and Blockhistogram-512. The results show that although the recognition rate is little bit lower but it is still close and allows us to see how much (or little) we stand to gain for the additional memory and computing cost. We use the compact representations in the following fusion experiments.

TABLE IV
COMPARISON BETWEEN HISTOGRAM FEATURE BEFORE AND AFTER COMPACTION

Inner lip texture feature	SMO	BayesNet	Average
Histogram-32	57.3	64.5	60.9
Blockhistogram-64	58.9	66.4	62.7
Histogram-256	59.2	67.5	63.4
Blockhistogram-512	**62.3**	**71.8**	67.1

Combining with the geometric feature together, we test the joint feature with each inner lip feature. The experiment results are shown in Table 5. From this table, we could find that recognition rate in the joint features with all feasible texture features are higher than that in geometric one only. It proves that effectiveness and importance of the inner lip texture in visual speech representation. The results in Table 5 also show that after combining with geometric feature, the recognition results improve comparing with each individual texture feature in representing the visual speech. Especially to the local inner texture features, such as the teeth proportion and color moment, the recognition results improve largely. It proves that the contribution of geometric and the local inner lip textures are complementary for visual speech recognition. It is helpful in improving the bi-model recognition results using the fusion method in much compact way. Looking at the holistic inner lip texture features in Table V, the block histogram has the highest recognition rate in both cases before and after feature fusion. Together with the geometric feature, the recognition results indeed improve, but comparing with that of local texture features such as teethe proportion or color moment, the increasing extent is not so large. This is to some extent reflecting that the redundancy information exists between the holistic inner lip texture feature and the geometric feature for Chinese viseme recognition. From this point, the local inner lip texture, teeth proportion has obvious advantage in computing cost with lower dimension and relatively higher recognition results when fusion with the lip shape feature together. It only has 3 extra dimensions adding to the geometric feature, but on average 18.5% increase in accuracy (26.3% using SMO and 11.1% using BayesNet) compared with using the geometric feature only.

We don't specifically discuss the performance of using the different classifiers for Chinese viseme recognition, but the experiment results in Table 3-5 still reflects the difference existing when using various classifiers. For example, with BayesNet using the local inner texture: Blockhistogram as the feature, it has the highest viseme recognition rate whilst SMO with DCT has the highest accuracy rate. Therefore how to employ the complementary contribution of classifiers such as multi-classifier fusion to improve the overall recognition performance remains a matter for future investigation.

TABLE V
RECOGNITION RATE OF JOINT FEATURE WITH EACH DIFFERENT INNER LIP TEXTURE FEATURE

individual Feature	SMO	BayesNet	Average
G	52.0	55.0	53.5
Teethproportion	42.5	39.8	41.2
Colormoment	39.2	36.9	38.1
Blockaverage	52.1	54.8	53.5
DCT	63.7	59.0	61.4
Histogram-32	57.3	64.5	60.9
Blockhistogram-64	58.9	66.4	62.7
Joint Feature	**SMO**	**BayesNet**	**Average**
G+Teethproportion	65.7	61.1	63.4
G+Colormoment	56.3	58.8	57.6
G+Blockaverage	56.3	57.6	57.0
G+DCT	64.6	60.2	62.4
G+Histogram-32	**67.6**	***69.2***	***68.4***
G+Blockhistogram-64	***67.3***	**69.9**	**68.6**

C. CHMM based Bimodal Chinese Character Speech Recognition System

We employ our proposed joint visual speech feature, viz. the local inner lip texture and the geometric feature

into the bimodal Chinese character speech recognition system to test its effectiveness further.

Overview of Bimodal Speech Recognition System: The system block diagram is shown in Figure 13. The speech video is chopped into the Chinese character clips and the audio and visual speech sequences are processed separately as input of Coupled Hidden Markov Model (CHMM) to do the fusion Chinese character recognition. Here we use the Mel-Frequence Cepstral coefficient (MFCC) as the audio feature and our proposed joint feature to represent the visual speech images.

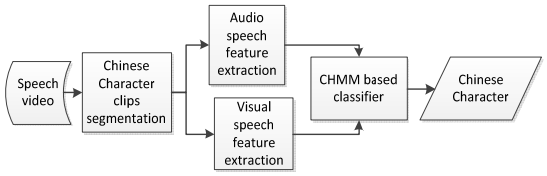

Figure 13. Block Diagram of CHMM-based Chinese Character Speech Recognition System

CHMM[19,20] is an effective solution to realize the medium fusion at the state level shown in Figure 14, where cycle symbols represent the hidden level, viz. states (the upper level represents the audio states and lower level is the visual states), and rectangle symbols represent the observed node, viz. audio and video inputs separately in the graph. This model takes the time correlation multiple modalities into account and solves the asynchrony problem of audio visual speech fusion.

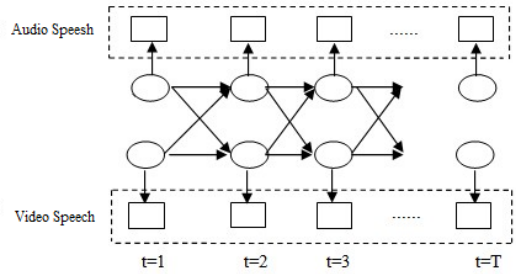

Figure 14. Structure of coupled HMM

In our CHMM based bimodal speech recognition system, we refer to the approach introduced in reference [21] to realize the training and recognition. The states and observers in the CHMM are combined and reconstructed into the structure like that in Figure 15 under the transition, where observe nodes use the different symbols to represent the corresponding compound observers. Then the traditional HMM training and testing algorithm such as Baum-Welch are used in the CHMM based bimodal speech recognition system.

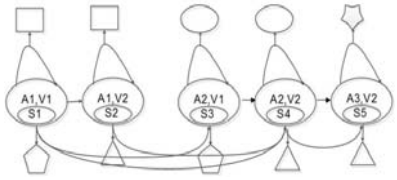

Figure 15. Illustration of equivalent transformation of CHMM

Bimodal Chinese Character Pronouncing Dataset: We test our approach on our own collection bimodal dataset. The corpus is selected based on Chinese phoneme defined in the SAPI[22], which actually points to the Pinyin of Chinese characters including initial consonant and vowel. We select 200 Chinese characters and 20 local Beijing mandarin speakers (ten females and ten males) to reading these Characters. We capture the frontal faces of these subjects using Sony HDR-XR160E and each character is repeated ten times. So there are 10 by 20 samples for each Chinese character.

In this pilot study, we only use the 10 Chinese characters listed in Table VI to do the experiment. The symbol column lists the used Chinese characters and example column is the reading words containing these Chinese characters marking with the tones. It covers most Chinese visemes described in section III. To separate the character pronouncing from the words not the single character has the natural pronouncing states and results are much practical.

TABLE VI
CHINESE CHARACTERS USED IN EXPERIMENT

Symbol	Example	Symbol	Example
ban	mu 4 ban 3	jiang	jiang 1 he 2
ci	ci 2 qi 4	mo	mo 4 shui 3
da	da 4 jia 1	qi	qi 3 qiu 2
deng	ban 3 deng 4	she	mang 3 she 2
gong	ban 3 deng 4	zhuang	mang 3 she 2

Experiment results and analysis: Taking the 60 samples of each Chinese character as training data and the other 40 as the testing samples, we do a series of comparative experiments with individual modality HMM and multimodality CHMM. Here we select 3-states/15-observers for audio-HMM and 3-states/ 9-observers for visual-HMM, and then 5 compound states for CHMM and coherent observers similar to HMM. K-means clustering algorithm is utilized in obtaining the observer serial number. Multi-observer based Baum-Welch training algorithm is used in the paper.

Each Chinese character's recognition results are listed in Table VII and Table VIII. Here 'A' in the Table represents the audio with MFCC as feature. The visual speech feature we test the geometric feature only and joint feature with several local inner lip texture features. The number indicates the number of characters (phonemes) correctly recognized out of 40 test samples (the result that is best within the visual modality with 2 errors or 5% is shown in italic; bold if comparable to or better than achieved with the other; chance level is 4 right or 10%).

Analyzing the above experiment results, we can find the results as follows.

Accuracy rates with CHMM based bimodal speech recognition system are much higher than that with HMM based single modal system. With CHMM, the average recognition rate of all feasible features is 93.4% and increases by 43.8% comparing with visual speech and 14.6% with single audio speech.

TABLE VII
HMM-BASED SINGLE-MODAL CHINESE CHARACTER SPEECH RECOGNITION RESULTS (CORRECT OUT OF 40 TEST CHARS)

Feature	ban	ci	da	deng	gong	jiang
A	**35**	**34**	**34**	**31**	*32*	**36**
G	23	6	19	8	25	12
G+Teethprotation	29	7	16	3	28	16
G+Colormoment	17	15	*27*	14	**35**	11
G+Blockaverage	**35**	**26**	*26*	12	*31*	7

mo	qi	she	zhuang	Total Correct	Accuracy (%)
22	**40**	**37**	14	**315**	**78.8**
29	3	9	27	161	40.3
37	16	19	*31*	202	50.5
28	12	*29*	**33**	*221*	*55.3*
24	12	4	28	205	51.3

TABLE VIII
CHMM-BASED BIMODAL-MODAL CHINESE CHARACTER SPEECH RECOGNITION RESULTS

Feature	ban	ci	da	deng	gong	jiang
G+A	*38*	*33*	*37*	**38**	*36*	36
G+Teethprotation+A	**40**	**34**	**39**	32	**38**	*38*
G+Colormoment +A	**40**	**34**	*38*	**38**	*37*	**40**
G+Blockaverage+A	*39*	**34**	36	35	*37*	36

mo	qi	she	zhuang	Total Right Number	RATE(%)
40	**39**	36	*37*	*370*	*92.5*
40	**39**	**39**	*36*	*375*	*93.8*
40	**39**	**39**	**38**	**383**	**95.8**
40	**39**	36	*36*	*368*	*92*

Under this situation, the joint feature geometric with the local color moment in the L color channel has the highest accuracy rate 55.3% in HMM system and 95.8% in CHMM system. It proves the effectiveness of the joint feature and in particular the utility of the local inner texture feature. Our results also show that for multiple subjects, the color moment feature shows stable performance in contrast with the teeth proportion apparently owing to individual differences in teeth appearance and habit. The local inner texture feature also has lower computational cost and lower dimensionality.

IV CONCLUSION AND FUTURE WORK

The paper proposes a joint visual speech feature which combines the geometric feature with several feasible features for describing the local inner lip texture. The feature are fused at the feature level and normalized with the simple min-max technique. We demonstrate the effectiveness of the joint feature through experiments on Chinese viseme recognition system and a bimodal Chinese character recognition system. The experiment demonstrates the potential of our joint visual speech feature, including especially the geometric and local inner texture feature, which has the good ability to represent speaking mouths no matter in in a multi-speaker bimodal phoneme plus viseme based Chinese speech recognition system. The complementary nature of the information in the bimodal features reveals the state of visual speech from different point of view. The joint features also have obvious advantage in terms of computational cost due to lower feature dimension required in the visual modality. The comparison experiments between HMM and CHMM-based recognition system shows that the CHMM system outperforms HMM based system with a considerably higher recognition rate of over 95% versus HMMs at under 80% with audio alone and at best around 55% for visual alone.

In the future work, we will explore the contribution of different classifiers and improve the performance visual speech recognition by employing multi-classifier fusion. AAM based feature point location needs to be discussed further to improve the automatic location accuracy to provide the basis for visual feature extraction to obtain more label data for further experiments to test the generation of our proposed joint feature.

ACKNOWLEDGMENT

This work was supported in part by a grant from the Chinese Natural Science Foundation under Grants No.61070117, the Beijing Natural Science Foundation under Grant No.4122004, the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions

REFERENCE

[1] Yaming Wang, Fuqian Tang and Junbao Zheng, "Robust Text-independent Speaker Identification in a Time-varying Noisy Environment," *Journal of Software,* vol. 7(9), 2012, pp.1975-1980.

[2] Neti, C.Potamianos, G.Luettin, "Large-vocabulary audio-visual speech recognition：a summary of the Johns Hopkins Summer 2000 Workshop," *Multimedia Signal Processing 2001 IEEE Fourth Workshop on.* 2001, pp. 619-624.

[3] Mehrabian A. "Communication without words," *Psychology Today*, vol. 2(4), 1968, pp.53-56.

[4] XinGuang Li, MinFeng Yao, JiaNeng Yang , "Speech Recognition Approach Based on Speech Feature Clustering and HMM," *Journal of Computers*, vol. 7(9), 2012, pp.2269-2276.

[5] Potamianos G, Neti C, Gravier G, et al. "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE,* vol. 91(9): 2003, pp.1306-1326.

[6] Wu Z, Aleksic P S, Katsaggelos A K. "Lip tracking for MPEG-4 facial animation," *Fourth IEEE International Conference on Multimodal Interfaces,* 2002, pp. 293-298.

[7] Stillittano S, Caplier A. "Inner lip segmentation by combining active contours and parametric models," *Proceedings of the 3rd International Conference on Computer Vision Theory and Applications (VISAPP'08)*. 2008, pp.297-304.

[8] Zhang X, Mersereau R M, Clements M, et al. "Visual speech feature extraction for improved speech recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, 2002, pp.1993-1996.

[9] Zhi Q, Cheok A D, Sengupta K, et al. "Audio-visual modeling for bimodal speech recognition," *2001 IEEE International Conference on Systems, Man, and Cybernetics,* 2001, pp.181-186.
[10] Zhi Q, Cheok A D, Sengupta K, et al. "Analysis of lip geometric features for audio-visual speech recognition," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans,* vol.34(4) , 2004, pp.564-570.
[11] Matthews I, Cootes T F, Bangham J A, et al. "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.24(2), 2002, pp.198-213.
[12] XinGuang Li, MinFeng Yao, JiaNeng Yang, "Feature extraction in speechreading," *Journal of Software*, Vol.5(7), 2010, pp. 705-712.
[13] Chan M T. "HMM-based audio-visual speech recognition integrating geometric and appearance-based visual features," *2001 IEEE Fourth Workshop on Multimedia Signal Processing*, 2001, pp.9-14.
[14] Benhaim, Eric,Sahbi, Hichem and Vitte, Guillaume, "Designing relevant features for visual speech recognition," *ICASSP 2013, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 2420-2424.
[15] Topkaya, Ibrahim Saygin, Erdogan, Hakan, "Using multiple visual tandem streams in audio-visual speech recognition," *ICASSP 2011, IEEE International Conference on Acoustics, Speech and Signal Processing,* 2011, pp. 4988-4991.
[16] Lewis, Trent W. and Powers, David M. W., "Audio-Visual Speech Recognition using Red Exclusion and Neural Networks," *Australasian Computer Science Conference (Sydney: ACS, 2002)*, 2002, pp.149-156.
[17] Boodoo N B, Subramanian R K. "Robust multi biometric recognition using face and ear images," *International Journal of Computer Science and Information Security*, vol. 6(2), 2009, pp.164-169.
[18] Zhao Hui, Chen Yue-Bing, Shen Ya-Min, Tang Chao-Jing. "Audio-visual speech synthesis based on Chinese visual triphone," *Proceedings of the 2009 2nd International Congress on Image and Signal Processing, CISP'09*.
[19] Rao B D, Trivedi M M. "Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition," *ICASSP 2008. IEEE International Conference on Acoustics, Speech and Signal Processing,* 2008, pp.2241-2244.
[20] M.Brand, "Coupled hidden Markov models for modeling interacting process," *tech.Rept TR 405, MIT Media Lab,* 1996, pp.150-156.
[21] Gang Liu, Wei Chen, Jun Guo, "A Research on Mixture Splitting for CHMM Based on DBC," *Journal of Computers*, Vol 4(11), 2009, pp.1167-1175.
http://msdn.microsoft.com/enus/library/ee125160(v=vs.85).aspx.

Xibin Jia, was born in Shanxi, China in 1969. Xibin Jia received PhD degree in computer science and technology from Beijing University of Technology in 2007, M.E. degree in intelligent instrument from North China Institute of Technology in 1996 and B.E. degree in wireless technology from Chongqing University in 1991. Xibin Jia was a visiting scholar at Flinders University, South Australia for 6 months from Sep. 2009 to Mar. 2010.

She is an Associate Professor in the College of Computer Science and Technology at the Beijing University of Technology in Beijing, China. Her recent publications include:
1. Xibin Jia, Yanfeng Sun, A Kind of Visual Speech Feature with the Geometric and Local Inner Texture Description, Telkomnika, 2013, Vol.11(2):877-889 .
2. Xibin Jia, Yanfang Han, David Powers, Spatial and temporal visual speech feature for Chinese phoneme, Journal of Information and Computational Science, 2012, Vol.9(14):4177-4185.
3. Xibin Jia, Xiyuan Bao, David Powers, Yujian Li, Facial Expression Recognition Based on Block Gabor Wavelet Fusion Feature, Journal of Convergence Information Technology, 2012, Vol.8(5):282-289.

Her areas of interest include visual information cognition, and multi-information fusion, especially for facial expression recognition and visual speech recognition.

Assoc.Prof. Jia is a member of IEEE, CCF.

Hua Du, was born in Inner Mongolia, China in 1989. He earned bachelor degree in computer science and technology from Beijing University of Technology in 2012. Hua Du focuses on visual information cognition, pattern recognition and machine learning.

He is a MS candidate in Computer college of BJUT in Beijing, China. His current publications include: Xibin Jia, Hua Du, Yanfang Han, Kewei Zhang, David Powers, "Audio/Visual Speech based Pronunciation Automatic Evaluation Algorithm and Comparison Platform", IJIIP: International Journal of Intelligent Information Processing, Vol. 4, No. 1, pp. 98 ~ 104, 2013.

Mr. Du is a member of CCF.

Yanfang Han, was born in Henan, China in 1988. She earned master degree in computer science and technology from Beijing University of Technology in 2013, bachelor degree in computer science and technology from Xinyang Normal University in 2010. Yanfang Han focuses on visual information cognition, image processing and pattern recognition.

Her current publications include:1. Xibin Jia, Yanfang Han, David Powers, Spatial and temporal visual speech feature for Chinese phonemes, Journal of Information and Computational Science, 2012, Vol.9(14):4177-4185. 2. Xibin Jia, Hua Du, Yanfang Han, Kewei Zhang, David Powers, "Audio/Visual Speech based Pronunciation Automatic Evaluation Algorithm and Comparison Platform", IJIIP: International Journal of Intelligent Information Processing, Vol. 4, No. 1, pp. 98 ~ 104, 2013.

David M. W. Powers was born in Sydney, Australia and earned a Bachelor of Science from the University of Sydney in 1979 with a major in pure mathematics and Honours in computer science including a focus on programming languages and artificial intelligence. His PhD was undertaken in the School of Electrical Engineering at the University of New South Wales in computational

psycholinguistics, as a pioneer in the new interdisciplinary fields of unsupervised learning, cognitive science and machine learning of natural language, the PhD being awarded in 1989 and published by Springer in the same year. He also has qualifications in linguistics (Summer Institute of Linguistics, 1980), theology (Moore Theological College, 1984), technical analysis of financial markets (Securities Institute Australia, 2001) and teaching English as a second or other language (LinguaEdge, 2011).

He is currently Full Professor of Cognitive and Computer Science, Associate Dean (International) and Director of the Centre of Knowledge and Interaction Technologies, in the School of Computer Science, Engineering and Mathematics, Flinders University, Adelaide, South Australia, as well as Visiting Professor at the Beijing University of Technology, with support from the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions. Previous positions were at Telecom Paris (1993), University of Tilburg (1992), University of Kaiserslautern/German Artificial Intelligence Institute/DFKI (1989-1991), Macquarie University (1984-1989) and University of New South Wales (1983-1984). Significant publications include

1. David M. W. Powers, *Parallelized QuickSort with Optimal Speedup*, in N.N. Mirenkov, Parallel Computing Technologies **(**Singapore**:** World Scientific, 1991): 167-176.
2. David M. W. Powers. *How far can self-organization go? Results in unsupervised language learning,* Spring Symposium on Machine Learning of Natural Language and Ontology (Stanford: AAAI, 1991): 131-136.
3. Jin Hu Huang, and David M. W. Powers (2001), *Large scale experiments on correction of confused words*. Australasian Computer Science Conference (Sydney: ACS, 2001): 77-82.
4. David M. W. Powers, *The Problem with Kappa*, Conference of the European Chapter of the Association for Computational Linguistics (Avignon: ACL, 2012): 343-355.

Prof. Powers is a life member of AAAI, a senior member of IEEE, and a member of ACM and ACL, previously holding the positions of Editor-in-Chief of ACM SIGART and Founding President of ACL SIGNLL, as well as a being a member of ACS SA Branch Executive and National Technical Committee.