

---

## A Kind of Visual Speech Feature with the Geometric and Local Inner Texture Description

Xibin Jia<sup>\*1</sup>, Yanfeng Sun<sup>1</sup>

Multimedia and Intelligent Software Technology, Beijing Municipal Key Laboratory Beijing University of Technology, Beijing, China

\*Corresponding author, e-mail: jiaxibin@bjut.edu.cn

### Abstract

*In this paper, we propose a type of joint feature with geometric parameters and color moments to represent the speaking-mouth frames for image-based visual speech synthesis systems. Based on FDP around the mouth area, the geometric feature is obtained by computing Euclidean distances to describe the width of the speaking mouth, the height of the outer and inner lips and the distances between them. The color moment component in the joint feature is obtained by calculating the texture between the upper and lower inner lips to describe the visibility state of the teeth. Through analyzing the accordance between the teeth visibility and the components of RGB and HSV color space based on the samples separately, we discovered that green and blue components are good at describing the change of teeth visibility. The experiments show that the proposed joint feature can effectively provide the basis for categorizing the different speaking states especially at the sense of lip shapes and tooth visibility. The evaluation of clustering results is done by analyzing the derived parameters of the silhouette function. The analyzing results prove that comparing with the geometric only and PCA, our proposed feature together with the shape and the local inner lip texture clues has better performance in improving the similarity between samples within the clusters. In the future, more expressive features with the shape and local texture information should be explored to increase the proportion of similar samples within the clusters to improve the descriptive ability of speaking mouths.*

**Keywords:** speaking-mouth image, visual speech feature, geometric feature, local texture of inner lip

Copyright © 2013 Universitas Ahmad Dahlan. All rights reserved.

### 1. Introduction

Humanoid figures can improve the friendliness of the human-machine interface, especially those figures that engage in speaking behavior. The synthetic results, often called talking head or visual speech, could soon be widely used in the area of education, electronic commerce, broadcasting or entertainment as a virtual agent, with applications such as virtual teacher, virtual seller or virtual desktop agent [1]. Pandzic at AT & T labs [2] showed that a talking head aids users' understanding of spoken text in noisy conditions (error rates drop from 16% to 8%), it can effectively provide a service for business and make waiting times more acceptable to the user, and a talking head makes services more attractive to the users, particularly when they compare directly the same service with or without the facial animation. Broadly, there are two categories of approach to synthesize a talking head: 3D model-based methods and image-based methods. The 3D model-based approach, including parameter models and muscle models, have achieved many excellent results [3, 4]. With the aid of a 3D scanner to obtain the visual speech shape and appearance information, the reality of synthetic results improves a lot; however, the robotistic results appear artificial, and are still easy to recognize. Part of the reason is due to the difficulties in emulating and modeling speaking behaviors such as the lip, jaw and tongue movements of a natural person. An image-based visual speech synthesis system has more realistic synthesis results, for it synchronizes the acoustic speech with the speaking mouth animation by re-ordering the pre-recorded speaking mouth image frames rather than establishing a complex mathematical model of the speaking action. The relationship between the acoustic speech feature and the corresponding mouth image feature is learned from video footage captured of a subject speaking [5-8]. From the viewpoint of improving reality, our system employs the latter method, aiming to obtain more realistic synthesis results and learn the potential relationship between acoustic speech and

visual speech, which we foresee can be used for driving the 3D model. The research of Brand [5] gives a good example of this application. In his voice puppetry system, Brand tracks 26 points on a face to describe the speaking action; this series of facial control parameters is suitable for driving many different kinds of animation ranging from video-realistic image warps to 3D cartoon characters.

Because the rationale of the image-based methods is to predict or synthesize visual speech from searching and reordering the captured clips, comprehending visual speech images is a key step. In other words, the speech animation information is contained in the existing visual speech images and is used to discriminate the speaking behavior states of these images. It is not difficult for a natural person to recognize the mouth moving states, but which parts of the whole face play more important roles in human perception, and the best means to extract the central feature and model are still topics under study. So how to get and describe the visual speech information from images is a major challenge to the image-based visual speech synthesis approach. From a psychological viewpoint, research on lip reading and speech reading provides us with some conception of how people perceive speech; as with the image based syntheses approach, speech reading research also depends on the comprehension of speaking captured on video. The difference that may exist between them lies in the emphasis of the visual speech synthesis system on determining the speaking state from the speaking mouth images rather than determining the meaning. So, in the image-based visual speech synthesis system, the effective representation of the speaking mouth image frames must be considered to get the correct understanding of the speaking states.

Generally, there are two kinds of approaches to describe the visual speech images: approaches based on global features include principal component analysis (PCA) [6] and local linear embedding (LLE) [7]; other approaches, based on local features, are reliant on geometric features [5, 18, 19]. PCA has been applied to reduce the dimension of images through establishing the eigenface or eigenlip images from analyzing the intensity statistic distribution of the sample images. Williams and Katsaggelos constructed 40 eigenlip images as the orthonormal basis for modeling the lips in their system, and 40 PCA coefficients were employed to represent the speaking mouth images, which represent 95.5% of the statistical variance [8]. Hancock et al. [9] used PCA in face processing to show that PCA of a set of face images does a good job of accounting for some aspects of human face perception from a psychological point of view. In [10] DCT and LDA were also used to represent the mouth images in compact forms.

Graf et al. [7] approached the speech synthesis problem by combining global features. They employed LLE to represent speaking images with as few as 12 LLE coefficients, possible because LLE analyzes local neighborhoods to determine how many dimensions are significant; in contrast, PCA captures the overall dimensionality of the data. In the LLE representation, images that look visually similar lie closely together. Geometric parameters (e.g. width and height of the mouth) were combined with LLE to represent the mouth images to compare the similarity.

Consensus has not been reached concerning which information, global or local, most significantly impact natural people in their understanding of visual speech and which local features play more important parts in the comprehension of visual speech. It has been argued that the global features contain more information while the local features lose the details which may help people to perceive visual speech. On the other hand, the global features are more often seen to suffer from variations in illumination. Moreover, from the viewpoint of the physiological mechanism of speech production, the main components in the vocal tract responsible for sound production are the velum, tongue, teeth, lips and jaw. Lucey [11] believes that for a lip reading system, the area around the mouth that contains these visible articulators should be extracted.

It is largely agreed that most information pertaining to visual speech stems from the areas around the lips, even though visual speech is located throughout the human face to some extent [12]. Brooke et al. [13] found that the visible articulators such as teeth and tongue improved the perception of vowels. Finn [14] found that for consonants the most important features were the size and shape of the lips. Massaro et al. [15] have suggested that degree of lip opening could be a critical visual feature which acts as input to the audiovisual integration process. Kaynak et al. [16] compared and analyzed the importance of geometric visual features in audio-visual recognition. Experimental results show that the geometric visual features, especially, the lip vertical aperture is the most relevant; and the visual feature vector formed by

vertical and horizontal lip apertures and the first-order derivative of the lip corner angle leads to the best recognition results. The experiments of Lawrence et al. [17] confirm that isolated kinematic properties of visible speech can provide information for lip reading. In their experiments, only the point-light (PL) locations on the lips, tongue and other parts of the face were displayed. Accordingly, on video tape filmed of the subject speaking, only the dots and their motions were visible. Although the static point light images can't be seen as faces, observers of the point-light stimuli describe the image as closely resembling a mouth. Test results for correct responses between the discrepant point-light condition and the audio alone condition indicate that the point-light stimuli helps observers to understanding the speech.

These types of research provide an explanation of how a natural person usually comes to understand visual speech from perceiving some local feature of the lip area such as the following geometric information: the lip contour, the extent of lip opening, rounding and protrusion. Accordingly, the local feature-based speaking-mouth image representation is widely used in image-based visual speech synthesis systems. Bregler et al. [6], in their video rewrite system, used 54 eigenpoints to find the mouth and jaw and to label their contours. They measured the similarity of the speaking mouth images by computing the Euclidean distance between four-element feature vectors containing the overall lip width, overall lip height, inner lip height, and height of visible teeth. Masatsune et al. [18] used ten position parameters to represent the lip shape, namely the vertical distance from the nose to the corner of the mouth, the horizontal opening of inner contour, and the vertical distances from horizontal axis, which is the line joining mouth corners to the inner contour at eight equally spaced points between the mouth corners. The geometric feature is good at describing the mouth shape, but it ignores texture clues such as the visibility of teeth, which can distinguish the different speaking states. Actually the appearance model including both the shape and the texture model is now widely used in face modeling. In reference, Cosker et al. [19] utilized the appearance feature in the video reality talking head synthesis system to establish hierarchical non-linear speech-appearance models.

Based on the above review of studies, our system concentrates on the lip area and aims to find a simple and effective way to represent the speaking-mouth image. Considering the aim of our system is to construct the mapping relationship at the feature level between the acoustic speech and the visual speech from the training data corpus, the proposed image feature should help the system to distinguish the different speaking mouth states corresponding with the pronouncing words. Therefore, taking account of both the shape and the texture clue of teeth together, we propose a joint feature with the geometric feature generated on the basis of the FDP around the lip region and the local texture feature reflecting the visibility of the teeth.

The remainder of this paper is organized as follows. In Section II, the overview of the system is given and the requirements of the speaking mouth images are analyzed in the context of the system. In Section III, we introduce the joint lip feature with geometric and color moment. In Section IV, the discriminative performance analysis of the corresponding feature on the basis of experiment results is introduced. Finally, the experiment results are analyzed and the possible future job is given.

## 2. System Overview

In order to explain our proposed representation method of the speaking mouth image, we firstly introduce our visual speech synthesis system (shown in Figure 1) and explicate the requirements of our system. The rationale of our system is to learn the acoustic-to-visual speech (A/V) mapping relationship based on the data-driven approach. Accordingly, the system is divided into two stages. At the training stage, we aim to establish a loose mapping relationship from the acoustic speech categories to the visual speech categories based on the assumption that the visual class has an identical structure with that of an acoustic class within the same data collection. The K-means method is used to cluster the acoustic speech; subsequently, the mapping of the visual speech categories is generated on the basis of the captured video dataset [20]. The visual speech relative to acoustic speech feature mode is learned based on a genetic algorithm to improve the cohesive strength of the mapping relationship. The evolutionary procedures are controlled by the computation of the similarity of the visual speech in the same mapping visual speech categories and the disparity between the different classes, further details of which were given in our previous work [21]. At the synthesis

stage, the input acoustic speech is processed and the corresponding visual speech cluster series is outputted based on the training A/V mapping mode.

After the feature mode extraction and A/V mapping has completed as presented above, the Viterbi algorithm is then used to search for the smooth lip image sequence from the generated visual speech category sequence. Figure 2 illustrates the solution procedure. Assuming the speech cluster index series is 3-1-11-11-1-3, the image cluster index is obtained at the same index and the corresponding candidate images in cluster are given according to the A/V mapping mode. The search processing is done on the basis of similarity estimation of the images in the neighboring clusters and the optimal lip image series is obtained to synthesize the final speech-synch visual speech. Therefore, the similarity of neighboring images is also needed to be estimated to get the smooth visual speech series. In this case, the effective feature of the visual speech is needed by our system, which can distinguish the difference between the different visual speech images.

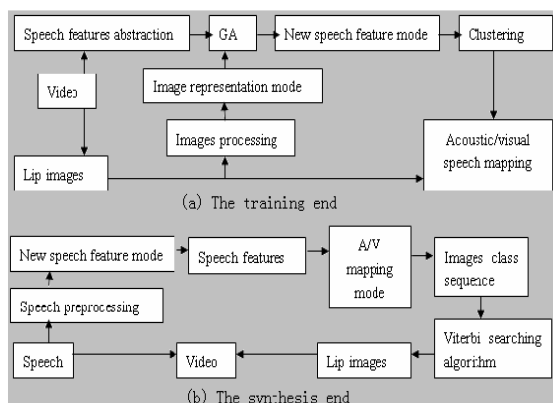


Figure 1. The Block Diagram of the Visual Speech Synthesis System

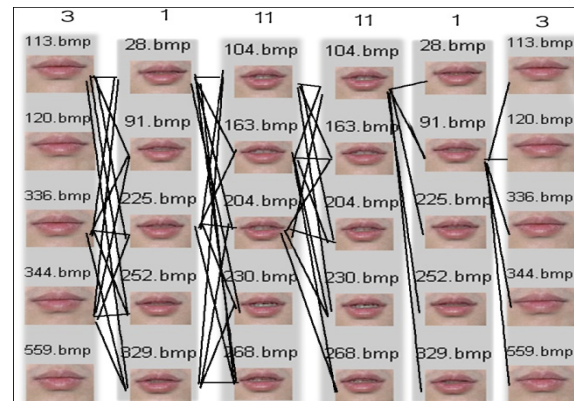


Figure 2. Illustration of the Rationale of Searching the Smooth Lip Image Sequence

In this work, visual speech refers to the images of the mouth area that are segmented from video footage filmed while the subject is speaking. The sentences are selected from a news corpus and are read by a male subject in standard Mandarin Chinese. The frontal view of the face is filmed during the reading, and the mouth area of each frame is partitioned at a fixed size and stored in the bitmap format. As the original values of these images in RGB color form do not include any specific meaning about the speaking states, we make some comparisons between PCA feature and geometric feature. Considering the importance of teeth visibility in identifying the speaking states, we evaluate the ability of the different texture forms and components in classifying the speaking images with different teeth visibility.

### 3. Joint Lip Feature with Geometric and Color Moment

#### 3.1. Representation of the Lip Shape

Analyzing the raw speaking mouth images and the character of the natural people perceiving the visual speech, the variety of the lip shape during speaking mainly lies in the change of the height and the width of the outer and inner lip relative to the opening or rounding extent of the mouth. Being an elastic object, the thickness of the lip also changes during speaking. So the feature that can include the above geometric parameters is needed here. Here, the paper adopts FDPs around the lip region defined in MPEG-4, shown in Figure 3(a), as the base and employs the Euclidian distance between every two points around the outer and inner contour of the lip to construct the geometric feature. This feature describes the height and width between upper and lower outer lip, between upper and lower inner lip, and also the thickness between inner and outer lip. In view of the symmetry of the lip, some repetitive components are deleted to reduce the feature dimensionality, and the final parameters

annotated with the line are shown in Figure3(b). Every Euclidian distance between every two feature points  $(x_i, y_i)$  and  $(x_j, y_j)$  around the outer or inner lip contour, is computed as (1).

$$r(t) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{1}$$

Finally, the geometric feature with 27 dimensions  $R=[r(1),r(2),\dots,r(27)]$ , is established to represent every visual speech image in the corpus.

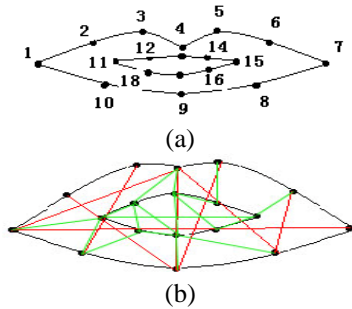


Figure 3. FDP and the Structure of Geometric Feature



Figure 4. The Images with Different Teeth Appearance



Figure 5. The Computation Area of Texture

**3.2. Representation of the Visibility of Teeth**

The geometric feature is good at distinguish the mouth shape, but some pronouncing states with similar geometric parameters, shown in Figure 4, represent different kinds of visual speech. Obviously the difference mainly lies in the state of teeth appearance. Actually, when a person is talking, the visibility of teeth relates to very different pronunciation, even though the shape is similar. To complement the incapability of the geometric feature in solving this problem, the feature to reflect the visibility of teeth is proposed in the paper.

The difference between the images with different teeth visibility extent mainly lies in the texture within the inner lip, especially the color clue. Therefore, the paper proposes to compute the first order central color moment  $C_k$ , shown in (2), to describe the texture inside the inner lip area. Here  $k$  represents the index of a processing image,  $l$  represents the number of components selected for determining the teeth visibility, as follows.

The corresponding components of the first order central color moment  $M_l$  are illustrated in (3), where  $t_l$  is the texture component of a pixel  $P_i$  in a certain color space. In RGB color space, for example,  $t_l$  represents the value of the red, green or blue component of pixel  $P_i$ , where  $l$  represents the relevant component.

$$C^k = \{M_1^k, M_2^k, \dots, M_l^k\} \tag{2}$$

$$M_l = \frac{1}{N} \sum_i^N t_l(P_i) \tag{3}$$

The sum of all  $P_i$  represents all pixels between the feature point 13 and 17 in Figure 3(a) and the other pixels along the three neighbor columns at the both sides of the pixels along the points 13 and 17. The area under computation is shown in Figure 5 with a cross symbol to mark each pixel. The neighboring pixels are computed to reduce the influence of the slot between the teeth. Here  $N$  is the number of pixels in this area.

To obtain a better understanding of which color form or components contribute to identification of visibility of teeth and their relative contributions, a consistency comparison is done between the color moment computed in RGB, HSV color space respectively and the corresponding states with different teeth visibility. Firstly, we exploit RGB color space to compute the color moment between the upper and lower inner lips. The means of the first order

color moments of the red (R), green (G) and blue (B) components are computed respectively according to (3). Figure 6 shows the results in bar graph form and the counterpart images which are selected from a continuous segment with typical speaking mouth states. We divide the graph into three parts with the yellow lines as label. From the speaking mouth images, we can tell they are three states with teeth not visible, partially visible and fully visible. The counterpart values of R, G and B locate at the different ranges which are in accordance with the above three parts judging from intuition. Moreover, if we observe the graph carefully, we can find that G and B components show more robustness whereas the R component is easily influenced by the lip texture, especially when the upper and lower lips are close. Like in part with teeth not visible and with the teeth partially visible, the values of the R color moment of some images are inseparable with the values located in the mixed ranges. The reason we can find in the paper of Lewis and Powers [22]. It exploits the red exclusion information to establish the feature of visual speech. Lewis and Powers point out that the rationale is that the face, including lips, are predominantly red, such that any contrast that may develop would be found in the green or blue color range (red exclusion). In this case, we prefer to use the G and B component to compute the texture between the upper and lower inner lips. Faridah et.al. [23] Uses the texture (energy, entropy, contrast, homogeneity) and color (R mean, G mean, and B mean) to perform feature extraction from parameters of coffee bean image samples together with the Neural Networks achieves good results of determining the extent of coffee bean quality. In view of relative experience, we make a quantitative analysis in the section IV, where we will compute the color moment (the mean of the color component) with R, G, B components and with G, B components only respectively together with the geometric feature to evaluate their ability to discriminate the speaking mouth states.

We also compute the first order color moment of the texture between the upper and lower lips in HSV space color, which separate the hue(H) information from the saturation(S) and value(V) information. In the same way as with the RGB color space, we make the comparison between the first order color moment value of hue, saturation and value components with the same images with the teeth visible to different extents. In the three different parts with three different teeth visibility states, the counterpart of the H, S and V first order color moment values are shown in Figure 7.

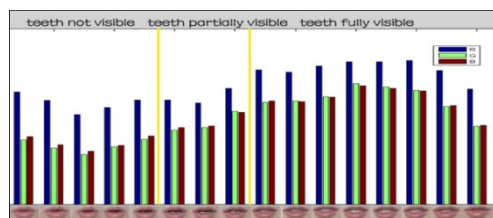


Figure 6. The Comparison Between the Different Teeth States and the Corresponding Color Moment of R, G, B Component

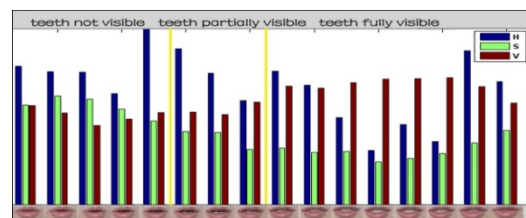


Figure 7. The Comparison Between the Different Teeth States and the Corresponding Color Moment of H, S, V Component

The results show that the color moment value computed with S component locates in a different range which is in accordance with the teeth visibility states. From the sense of intuition, it seems to have a better ability to reveal the teeth visibility states than the color moment derived from H and V components.

From the above analysis, we can see the first order color moment can be used to represent the states of teeth visibility, and the components of R and G in the RGB color space and the S component in the HSV color space are better candidates for describing the texture inside the inner lips. Further experiments will do to determine the final form. However, it is still not enough to describe the different teeth states when the upper and lower lips are close, and the inside of the mouth is dark and teeth visibility states cannot be distinguished very easy.

### 3.2. Joint Feature of the Speaking-mouth Image

The final feature proposed to be employed in the paper is a combination of the above geometric components and the color moment components. As these two components are

computed in different scale systems, the large parameters will dominate over the function of the small parameters. In order to balance the dedication weight of two kinds of feature in representing the speaking-mouth images, the balance coefficient is added, illustrated in (4). The final feature  $F^k$  is represented in a vector with two parts: the geometric component  $R^k$  and the color moment  $C^k$ , here  $k$  is the number of image frames in the footage and  $S$  denotes the sample space.

The balance coefficients  $\alpha$  and  $\beta$  are computed as shown in (5). Here the function  $\max()$  is employed to obtain the maximum value of the geometric feature and the color moment feature respectively and normalize the value of every component within the range of 0 to 1.

$$F^k_l = [a \times R^k \quad \beta \times C^k] \tag{4}$$

$$\alpha = \frac{\max_{k \in S} \{C^k\}}{\max_{k \in S} \{C^k\} + \max_{k \in S} \{R^k\}}, \beta = \frac{\max_{k \in S} \{R^k\}}{\max_{k \in S} \{C^k\} + \max_{k \in S} \{R^k\}} \tag{5}$$

**4. Experimental results of joint feature discriminative analysis**

A data corpus with 1000 images captured of the speaking subject was pre-processed and each image was represented by the joint feature according to formula (4) with the geometric and color moment of R, G, and B, shown in (1) and (2) respectively. The sentences were selected from news paragraphs so as to try to include all possible visemes. Considering our visual speech synthesis approach is to establish the acoustic to visual speech relationship at the feature level, ensuring that the data set includes all the possible visual speech states is more important. Therefore we use the K-means unsupervised clustering algorithm to cluster the images in the data corpus into 15 categories. The clustering results with the representative image and the corresponding sample number of each category are shown in Figure 8. From the results, we can tell from the angle of data distribution that the selected data has enough various speaking states and a reasonable number of samples. The category number is close to the number that the present literature uses to describe the Chinese visemes. The MPEG-4 defines 14 types of the viseme presented in [24] and Yan Jie [25] classifies the Chinese visemes into six kinds. So it also gives a proof from the clustering result that our joint feature can distinguish the different kinds of visual speech well.
















Classes					
Sample number	63	84	67	41	84
Classes					
Sample number	41	79	88	66	119
Classes					
Sample number	77	31	82	25	68

Figure 8. The Distribution of the Training Data

Table 1. Comparison of the Derived Silhouette Parameters Using the Different Image Features

Features parameters	PCA	Geometric feature only	Joint feature with geometric and color moment
	Mean	0.2646	0.3591
Maximum	0.8166	0.7854	0.9434
Percentage with a value greater than 0.5	14%	25.3%	38.6%

**4.1. Discriminative Analysis of Speaking Mouth Representation in Joint Feature and Geometric Feature Only and with PCA**

There are a number of different ways to represent the visual speech images. As we introduced before, two families of features comprise the general viewpoint. As a typical approach with global features, principal component analysis (PCA) is employed on our data corpus to represent the speaking mouth images. Whilst the geometric feature is used only as a kind of typical local feature, the joint feature, with geometric and teeth texture information computed in RGB color space, is utilized to represent the speaking mouth images in our data set.

As no unique geometric feature has been isolated that is superior for discrimination, we adopt our geometric feature mode to represent the visual speech images, and then provide a comparison with feature methods, including PCA and joint features. The data represented in different features are clustered in accordance with the K-mean algorithm. The clustering results are analyzed to judge the discriminative ability of these features. Analysis is done from two angles: (1) quantitative analysis, in which we employ some parameters to evaluate the clustering performance; and (2) qualitative analysis, to evaluate the perception of the clustering results.

To determine the quantitative evaluation of the performance of the representation features from the view of the within and between clusters, we use the silhouette function to describe the clustering results. The silhouette function  $S(i)$  is defined as in (6), as follows.

$$S(i) = \frac{\min(\text{AVGD\_BETWEEN}(i,k)) - \text{AVGD\_WITHIN}(i)}{\max(\text{AVGD\_WITHIN}(i), \min(\text{AVGD\_BETWEEN}(i,k)))} \quad (6)$$

Here,  $i$  designates the index of the speaking mouth image in a certain cluster from the K-mean algorithm, where the  $k$  in the formula represents the index of images in all other clusters except the cluster image to which  $i$  belongs. The function  $\text{AVGD\_BETWEEN}(i, k)$  is used to compute the average distance between the images  $i$  and  $k$ , and the images are represented with the corresponding feature. The function  $\text{AVGD\_WITHIN}()$  is used to compute the mean distance between the images represented in the corresponding feature mode within the cluster to which image  $i$  itself belongs. The distance here employs the square Euclidean distance. In the formula,  $\min()$  and  $\max()$  are two functions to compute the minimum and maximum value.

The silhouette value for each point is a measure of how similar that point is to other points in that point's cluster versus points in other clusters, and ranges from -1 to +1. By this definition, the larger the value of the silhouette, the better the clustering results for the point. The rationale for this is that when the distance between clusters is bigger and the points within each cluster are more immediate, the clustering result provides more descriptive information about the points in the region in terms of the speech state identification task.

In order to clarify the above idea and make the comparison more precisely, we calculate the following derived parameters of the silhouette: the mean, the maximum and the percentage with the value of silhouette is bigger than 0.5. The corresponding values are listed in Table 1 with three different features respectively. With the joint feature to represent the speaking mouth images, the average value of the silhouette is 0.4115 and 38.6% of image silhouette values are greater than 0.5; compared to 0.3591 and 25.3% in geometric feature only; 0.2646 and 45% in PCA. These numbers clearly indicate that using the joint feature to represent the speaking mouth images has a stronger clustering capability than using the geometric feature or PCA only. Meanwhile, using the geometric feature only is better than using PCA. In other words, using the joint feature, the different speaking states can be classified in different clusters and the similar speaking states can be identified. That's what we expected. But, back to the Figure 9, the silhouette of some values is too low, some of them are even negative and the percentage of the silhouette with high values is not ideal. This means that some of the images in joint feature are still too close between clusters and will lead to some errors in clustering. The feature form to represent the speaking mouth images still needs to be improved to enlarge the dissimilarity between clusters

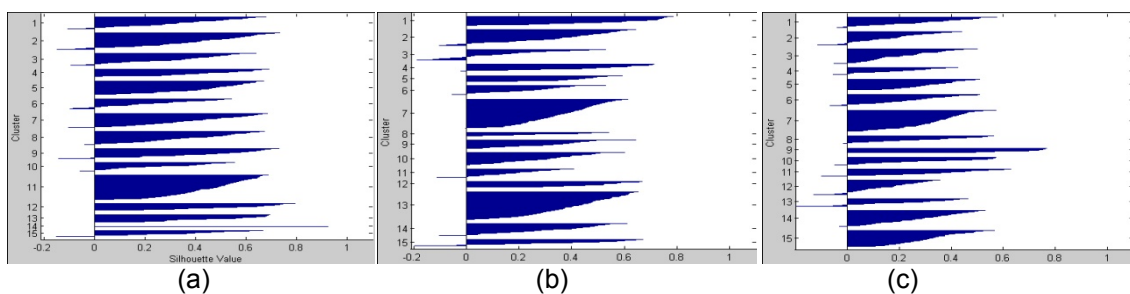


Figure 9. The Silhouette Graph with Speaking Mouth Represented in Various Features



The joint feature in form of the color moment to describe the teeth visibility states together the geometric feature is better than the feature with the geometric feature only. This can be proved by comparing the clustering deviation (also know as cluster error) from the centroid which is computed according to (7).

Here,  $X$  represents the image feature vector and  $m_i$  denotes the centroid of  $i$ th cluster and  $D_i$  represents the set of this cluster. It is computed according to (8) and  $n_i$  is the number of images within the  $i$ th cluster.

$$J_e = \sum_{i=1}^C \sum_{X \in D_i} \|X - m_i\|^2 \tag{7}$$

$$m_i = \frac{1}{n_i} \sum_{X \in D_i} X \tag{8}$$

With the joint feature and the geometric feature to represent the speaking mouth images respectively, the corresponding average clustering deviation  $J_e$  is 0.467952 and 0.498449. The two values illustrate that using the joint feature to represent the speaking mouth images produces less error than with the geometric feature only. In this case, the joint feature shows better clustering performance than the geometric feature without the teeth texture information from the point of view of evaluation of the within-cluster attribute.

In our image-based visual speech system, the final synthesis results largely depend on clustering results and another way to evaluate the clustering results is using visual inspection. In other words, the images within the same cluster should look the same and images from different clusters should look different. In order to make direct evaluation from the view of perception, the images clustered in the same cluster are displayed at same time and we make the intuitive estimate.

Firstly, we display the clustering results with the joint feature. Figure 10 shows the parts of the clustering results from the 15 clusters. Each block represents one cluster and shows the correlative images. To show that the joint feature has better ability to discriminate images with different teeth textures inside the inner lips, the 6 clusters are selected in pairs to show the contrast between clusters with different teeth visibility and similar lip shapes. The contrast pairs in Figure 10 are (a) vs (b), (c) vs (d), (e) vs (f).

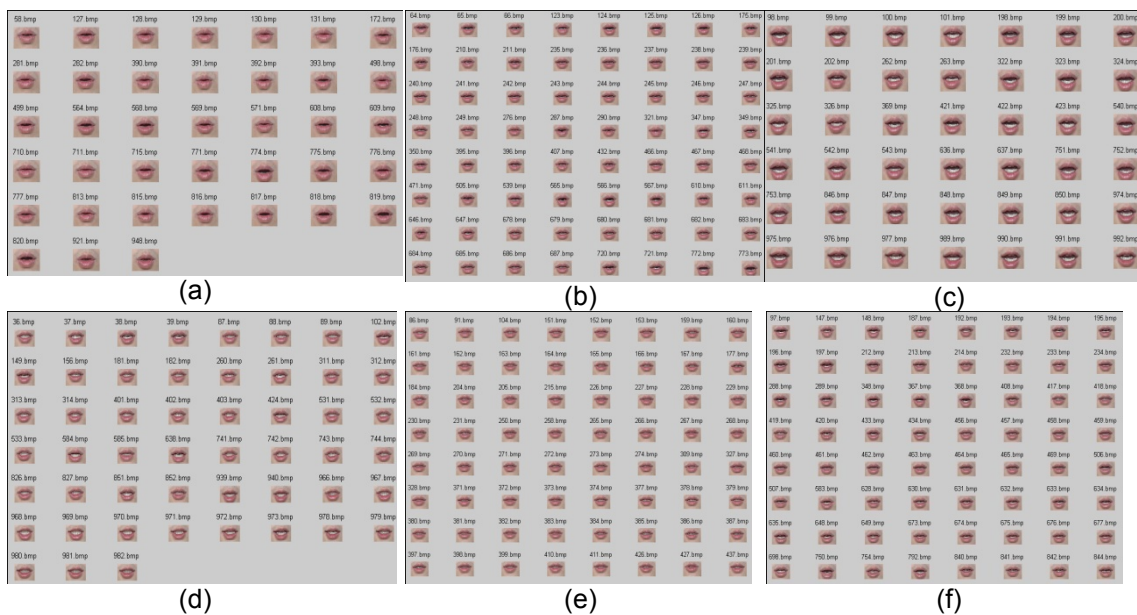


Figure 10. The Images in Part Clusters Based on Our Joint Feature

The clustering results are generated automatically with the K-means clustering algorithm. Comparing the images in different clusters, we can tell intuitively that the clusters represent the different mouth states: with the round shape and teeth invisible in Figure 10(a); with the round shape and teeth partially visible in Figure 10(b); with the big open mouth and teeth partially visible in Figure 10(c); with the big open mouth and teeth fully visible in Figure 10(d); and with mouth half open and teeth fully and partially visible in Figure 10(e) and (f) respectively. So the joint feature taking account of the geometric information and the texture information inside the inner lip can distinguish the states with different speaking mouth shape and teeth visibility to some extents. Of course, there still are some unexpected speaking mouth images being classified in a certain cluster.

For example in Figure 10(f), when the mouth half opens and teeth texture is not contrasted from the neighboring area largely with the shadow and the texture information can't tell the teeth visible states very well influenced by the change of illumination. So the feature mode still needs to be discussed to improve the discriminative ability of the speaking states.

To make an intuitive estimate of discriminative ability of the geometric feature only for representing the speaking mouth images, the same clustering processing is done on the same data set. Figure 11 shows part of the clustering results. As before, we select three clusters with round mouth, half opening and largely opening mouth shapes from the same view angle as in the above procedure. From the figure, we can tell that the images with different teeth visibility are mixed up in the same cluster. For instance, in the images in Figure 11(b), images with teeth fully visible are classified in the same cluster with teeth partially visible.

The clustering processing is done without involving any information of the teeth state. Apparently, the images classified in different clusters when represented with the joint feature are classified in the same cluster with geometric feature only. Some images with teeth opening-closing in the separate clusters shown in Figure 10 (c and d) are now in the same clusters in Figure 11 (b). Hence, texture information of teeth states can help to distinguish the various speaking states better. To identify these states is very important to the synthesis results. It is easy to perceive that with different teeth visible images mixed together to generate visual speech animation, even with a smooth mouth shape, the result is unacceptable.

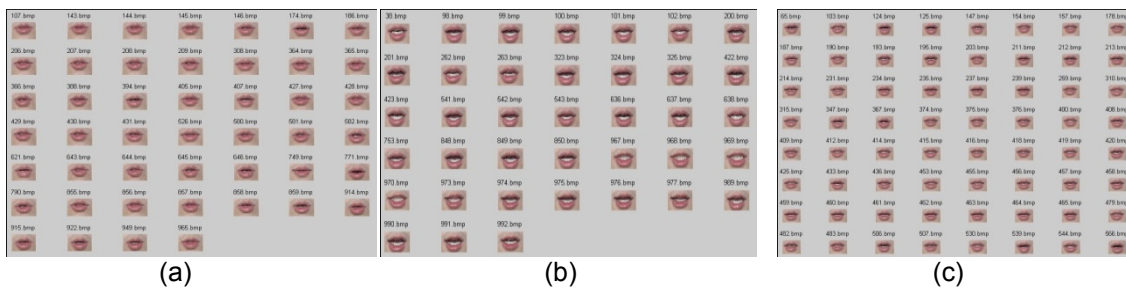


Figure 11. The Images in Part Clusters Based on Geometric Feature Only



Figure 12. The Images in Part Clusters Based on PCA

With PCA as feature to represent the speaking mouth images in same data set, the clustering results are unacceptable from human perception. Here we select 40 dimension PCA

coefficients. The images in two clusters are selected randomly from the 15 categories as shown in Figure 12. The obvious mixture in same clusters with the different speaking mouth images proves that PCA fails to serve as feature in this case. Part of the reason is due to the fact that the speaking-mouth image series are not very much different in intensity distribution, so it is hard to distinguish between them based on the statistic eigenlip.

#### 4.2. Discriminative Analysis of Teeth Texture Representation Method

From the above experiments, we can conclude that with the teeth texture information added to represent the speaking mouth images, the joint feature shows better discriminative capability for speaking states. However, because texture is unstable and is easily influenced by illumination, like the shadow when the upper and lower lips are close, we also carried out some other experiments about the representation of the teeth visibility.

According to the analysis about the accordance of the red, green, blue components in RGB color space and the hue, saturation, value components in HSV color space with the state of teeth visibility in section III, we find that the G and B components in RGB color space and S component in HSV color space display better coherence tendency with teeth visible states. Based on the above components in RGB and HSV color space respectively, the various feature modes combining the different possible components are used here together with the geometric feature to represent the speaking mouth images in data set. The following form is constructed to compute the first order color moment to represent the visibility of teeth between the upper lip and the lower lip according to (1), the components to combine the features are listed in Table 2. The first feature to describe the teeth texture is the one we used above in the joint feature: the first order color moment of whole R, G and B components. {R,G,B} represents the selected components. The second feature is derived from the G and B components, noted as {G, B} according to the idea that these two components are robust from the red skin background. The third feature is established through computing the color moment of the whole H, S and V components and other two features derived from the S and H component respectively are also shown here to figure out their attribute to identify the teeth states. The sixth color moment is computed based on the mixture of G, B and S components together to tell if the combination of the three better components can derive a better feature mode. Apart from these six modes derived based on the color space components, the grey scale intensity information is considered to describe the teeth visible states. And considering that the change of teeth texture illumination may contribute more strongly in identifying the teeth states, the gradient of the illumination is calculated as color moment.

Table 2. The Form of Feature to Represent Teeth Visibility

Index	The texture feature form
1	{R, G, B}
2	{G,B}
3	{H,S,V}
4	{H}
5	{S}
6	{G,B,S}
7	{Grey}
8	{Grey Gradient}



Figure 13. The Synthesis Sequence of "jianyi" in Pinyin

Table 3. Comparison of Silhouette Values in Different Texture Feature Modes

feature	Cluster number	Mean value of silhouette	Max value of silhouette	Percentage of the silhouette Value bigger than 0.5
{G,B}	15	0.4142	0.9547	39.7%
	6	0.5047	0.8426	55.3%
{R,G,B}	15	0.4115	0.9434	38.6%
	6	0.4364	0.8041	44.2%
{S}	15	0.4026	0.9462	36.9%
	6	0.4826	0.7962	53.7%
{G,B,S}	15	0.3988	0.9464	36.6%
	6	0.4928	0.9508	56.7%
{H,S,V}	15	0.3777	0.9191	32.1%
	6	0.4507	0.8205	47.8%
{H}	15	0.3796	0.8057	31.4%
	6	0.4640	0.8870	49.8%
{Grey}	15	0.3235	0.7368	31.4%
	6	0.4252	0.7807	40.2%
{Grey gradient}	15	0.3575	0.8608	24.1%
	6	0.4154	0.8007	40.6%

All these components are computed according to (2) and together with the geometric feature and the balance coefficient for normalizing each component according to the formula

(4), the joint features are computed to represent all images in the dataset. Employing the derived parameters of silhouette to evaluate their discriminative ability in classifying the speaking speech images, the corresponding results are shown in Table 3. To make the result more general, we cluster the data twice with 15 clusters and 6 clusters respectively. From the results we can determine which feature mode displays better performance in representing teeth visible states. The feature names in the Table only label the texture feature mode to represent the joint feature with corresponding color moment component and the geometric component together. The feature is ordered according to the mean of silhouette with 15 clusters to highlight the contribution capability of the following features. We can tell from the results in Table 3 that the texture component calculated on the G and B components in RGB color space displays higher discriminative ability than the others (with both 15 clusters and 6 clusters) than the others. Using the S component in HSV color space and G, B, S combined components to represent the texture between the upper and lower lips is better than using the hue component. The grey component with the lower silhouette value than the others proves the influence of illumination is stronger at this mode. These results are in accordance with the analysis of consistency comparison between the color space components and the change of the teeth visibility in section 3. From the experimental results, we confirm that green and blue components have better performance in describing the teeth texture information. The final texture component in our paper is determined by computing the first order center color moment of green and blue components together with the geometric feature to represent the speaking mouth images in our dataset.

## 5. Conclusion

In this section, it is explained the results of research and at the same time is given the compre This paper proposes a method to distinguish speaking mouth images by a joint visual speech feature with the geometric and first order central color moment of the local inner lip texture according to the understanding of ways natural persons to perceive the visual speech. The geometric feature is used to describe the lip shape. It has better category results with higher similarity between samples within the clusters than PCA in representing the lip shapes based on evaluating with silhouette and subjective analyzing of images in the same category. Together with the local inner lip texture, the clustering ability improves under the same experiments. Here the color moment parameter is used especially to emphasize the visibility of teeth, which reflects the different speaking state under the situation that the lip shapes are similar. The experiments show that using green and blue components to describe the local inner texture have the higher ability in discriminating the teeth visibility states than using a holistic feature such as PCA. Applying our proposed feature to represent the speaking mouth images, the system can obtain the right understanding of speaking images and have reasonable synthesis results. Figure 13 shows an example of the synthesized lip moving of “jianyi” in pinyin with the acceptable teeth visibility states.

Although the similarity within the same categories is improved with our joint feature in representing the speaking mouths, the proportion of samples with higher silhouette value is still less than expected and the relevance of features needs to be reduced. In future, exploitation on how to reveal the inner lip texture and establish an effective feature still needs to be done.

## References

- [1] Basori, Ahmad Hoirul, Tenriawaru, Andi Mansur, Andi Besse Firdausiah. Intelligent avatar on E-learning using facial expression and haptic. *Telkomnika*. 2011; 9(1): 115-124.
- [2] I Pandzic, J Ostermann, D Millen, User evaluation: Synthetic talking faces for interactive services, *The Visual Computer*. 1999; 15(7/8): 330-340.
- [3] Massaro DW, Ouni S, Cohen MM, Clark R. A Multilingual Embodied Conversational Agent. *Proceedings of 38th Annual Hawaii International Conference on System Sciences*. Los Alimitos, CA: IEEE Computer Society Press. 2005; 9: 296b-296b.
- [4] JP Lewis, Parke F, Automated lip-synch and speech synthesis for character animation, *CHI/GI 1987 conference proceedings on Human factors in computing systems and graphics interface*, Toronto, Ontario, Canada. 1987: 143-147.
- [5] M Brand. Voice puppetry, *Proceedings of ACM SIGGRAPH 1999*, ACM Press/Addison-Wesley Publishing Co, 1999; 21-28.

- [6] C Bregler, M Covell, Slaney, Video rewrite: Driving visual speech with audio, *Proc. SIGGRAPH'97, ACM*, Los Angeles, CA, USA. 1997; 353–360.
- [7] HP Graf, E Cosatto. Sample-based synthesis of talking-heads, *The 8th IEEE Int'l Conf. Computer Vision*, IEEE Comput. Soc, Los Alamitos, CA, USA, 2001; 3-7.
- [8] J Williams, K Katsaggelos, An HMM-Based Speech-to-Video Synthesizer, *IEEE Transactions On Neural Networks*. Institute of Electrical and Electronics Engineers Inc. 2002; 13(4); 900-916.
- [9] Peter J, B Hancock, A Mike Burton, Vicki Bruce, Face processing: human perception and principle components analysis , *Mem Cognit*. 1996; 24(1): 21-40.
- [10] G Potamianos, C Neti, Recent advances in the automatic recognition of audio-visual speech, *Proceedings of the IEEE*. 2003; 91(9): 1306-1326.
- [11] J Lucey. Lipreading across multiple views, Queensland university of technology, Ph.D Thesis, Brisbane, Queensland, Australia. 2007.
- [12] F Lavagetto. Converting speech into lip movements: a multimedia telephone for hard hearing people, *IEEE Transactions on Rehabilitation Engineering*. 1995; 3(1): 90-102.
- [13] N Brooke, A Summerfield. Analysis, sythesis, and perception of visible articulatory movements. *Journal of phonetics*. 1983; 63-76.
- [14] K Finn. An investigation of visible lip information to be used in automated speech recognition, PhD thesis. Georgetown University. Washington DC, USA, 1986.
- [15] DW Massaro, MM Cohen. Perception of synthesized audible and visible speech. *Psychology*. 1990; 1: 55-63.
- [16] MN Kaynak, Qi Zhi, AD Cheok, K Sengupta, Zhang Jian, Ko Chi Chung. Analysis of lip geometric features for audio-visual speech recognition. *IEEE Transactions on Systems Man and Cybernetics, Part A, Systems and Humans*. 2004; 34(4): 564 -570.
- [17] Rosenblum LD, Sadana HM, An audiovisual test of kinematic primitives for visual speech perception, *Journal of Experimental Psychology: Human Perception and Performance*. 1996; 22(2): 318-331.
- [18] Masatsune T, Shigekazu et al. Text-To-Audio-Visual Speech Synthesis Based On Parameter Generation From HMM. *Sixth European Conference on Speech Communication and Technology*, Budapest, Hungary. 1999; 959-962
- [19] D Cosker, D Marshall, P Rosin, YL Hicks. Video Realistic Talking Heads Using Hierarchical Non-Linear Speech-Appearance Models. *Proceedings Of Mirage 2003*. Inria Rocquencourt, France. 2003; 10-11: 22-27.
- [20] Xibin Jia, Baocai Yin, Yanfeng Sun, Xianping Lin, Learning based Visual Speech Synthesis System, *Journal of Information & Computational Science*. 2006; 3(2): 227-234.
- [21] Xibin Jia, Baocai Yin, Yanfeng Sun, Xianping Lin. GA-Based Speaking Mouth Correlative Speech Feature Abstraction, *ICCI2006 The 5th IEEE International Conference on Cognitive Informatics*. 2006; 114-119.
- [22] TW Lewis, DMW. Powers, Audio-Visual Speech Recognition using Red Exclusion and Neural Networks. *Journal of Research and Practice in Information Technology*. 2003; 35(1): 41-64.
- [23] Faridah, Parikesit, Gea OF, Ferdiansjah. Coffee bean grade determination based on image parameter, *Telkornika*. 2011; 9(3): 547-554.
- [24] Chen YQ, Gao W, Wang ZQ, Jiang DL. A speech driven face animation system based on machine learning. *Journal of Software*. 2003; 14(2):15-221.
- [25] Yan Jie, Text-Driven Lip Motion Synthesis System. *Computer Engineering and Design*. 1998; 19(1): 31-34.