Journal of Information Assurance and Security.
ISSN 1554-1010 Volume 9 (2014) pp. 148-156
© MIR Labs, www.mirlabs.net/jias/index.html

Persian Viseme Classification Using Interlaced Derivative Patterns and Support Vector Machine

Mohammad Mahdi Dehshibi[1], Jamshid Shanbehzadeh[2]

[1] Digital Signal Processing Lab., Pattern Research Center,
Karaj, Iran
info@dehshibi.com

[2] Department of Computer Engineering, Kharazmi University,
Tehran, Iran
jamshid@khu.ac.ir

***Abstract*: Viseme (Visual Phoneme) classification and analysis in every language are among the most important preliminaries for conducting various multimedia researches such as talking head, lip reading, lip synchronization, and computer assisted pronunciation training applications. With respect to the fact that analyzing visemes is a language dependent process, we concentrated our research on Persian language, which indeed has suffered from the lack of such study. To this end, we proposed an image-based approach which consists of four main steps, including (i) extracting the lip region, (ii) extracting Interlaced Derivative Patterns (IDP) considering coarticulation effect, (iii) using a hierarchical approach for clustering visemes in the Persian language by mapping each viseme into its subspace, and finally (iv) applying a Support Vector Machine (SVM) to classify visemes which their classes have been obtained in the previous step. In order to clustering visemes, we applied unweighted pair group method with arithmetic mean to each feature vector. Then, furthest neighbor of the weight value as a result of reconstruction is set as a criterion for comparing viseme dissimilarity in order to find appropriate clusters. Afterwards, obtained clusters have been considered as the classes to which phonemes should be classified. In order to indicate the robustness of the proposed algorithm, a set of experiments was conducted on AVA in which two syllables were examined. Comparing the results of the clustering and classification algorithms, regarding the extracted features, with that of the perceptual test given by an expert proves a reasonable evaluation of the proposed algorithms.**

Keywords: Audio/Visual processing, Persian Viseme clustering, Persian Viseme classification, Interlaced Derivative Patterns, Phoneme manifold.

I. Introduction

Various applications have been proposed in the field of audio/visual signal processing [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13] through the last forty years or so, among which are talking head [14], lip reading [15], lip synchronizing [16], computer assisted pronunciation training [17], and viseme clustering [18], [3]. Researches which are done in the area of viseme clustering are few [18], [19], although this phenomenon can produce appropriate audio/visual applications.

Visemes may be analyzed in two major ways. First but not foremost, humans can be employed to analyze the data, where some people with different lip reading abilities are taken to observe a syllable-sequence of meaningful and meaningless images and recognizes the uttered materials. Then, the overall results may be taken as the basis for an analysis of the visemes. Nonetheless, it alone cannot prove a perfectly reliable and accurate one, let alone of being time and cost consuming. Tony Ezzat [20] classified visemes based on subjective tests for 2D visual speech synthesis, where merely one word from several speakers is studied for each viseme. In the stated work, the coarticulation effect is excluded.

Second comes computerized algorithmic processing. In [18], an active contour model for inside mouth is considered to extract, which can be pursued in various utterances where one can trace points of interest in different utterances to detect lip variation, but it has neglected anterior movements of the lips in its clustering phase. Besides, it has low accuracy and high execution time while the iteration of this algorithm may cause local minimums. Krnoul et al. [21] have reached acceptable results in viseme similarity recognition and analysis for facial speech synthesis using PCA, through preparing good studio condition, and using infrared radiation and some reflective markers around lips; but it has ignored coarticulation. This method is not publically applicable, for the stated equipment used in it. In [22], Eigenspace with Bhattacharyya distance is used to measure visemes' similarity. In this work, utterances are pronounced in Spanish within 12 sentences, where merely continuous speech is considered. However, discrete syllables and the effect of coarticulation on visemes are overlooked.

Viseme grouping has also been carried out in Swedish language in [23], using the maximum likelihood classifier method aimed at sound phoneme articulation, and helping visual information and decreasing errors in Swedish language pronunciation. It has taken coarticulation effect into consideration, and it used video sequences recorded from one woman. In [14], a Persian talking head application is developed, where English phonemes and visemes are used instead of Persian ones, for Persian visemes were neither identified nor classified yet. This causes such products not to be photorealistic. Aghaahmadi et al. [19], cluster the Persian language visemes for the first time by the proposed novel and accurate algorithm, with respect to speech therapy applications and photo realistic talking head animation in target. Moreover, coarticulation and phoneme position in syllables are considered. Two female respondents participated in the study; the first one who was aware of sound speech rules was used in viseme clustering. They reduced the dimensions of training data by calculating Eigenlips for each of the visemes through Eigen analysis. Then the weight criterion out of the reconstruction of each viseme with the other is used for quantifying visemes' similarity. In [3], a hierarchical approach is used for clustering visemes in Persian language based on principal component analysis of a polynomial kernel matrix considering coarticulation effect. Having obtained feature vector of each phoneme, they applied unweighted pair group method with arithmetic mean to each projected viseme on the constructed manifold. Then a furthest neighbor of the weight value as a result of reconstruction is set as the criterion for comparing viseme dissimilarity. In order to indicate the robustness of the proposed algorithm, a set of experiments was conducted on Persian databases. Comparing the results of the clustering algorithm with that of the perceptual test given by an expert proves a reasonable evaluation of the proposed algorithm. Presently, over 150 million people in the world speak in Persian; nonetheless, there has regrettably been few such researches in this language, and clustering and classification viseme are of great importance. Viseme is the visual form of a phoneme [24]. In other words, visemes of some phonemes are alike, as /b/ and /p/ which are phonemically different, but the same in visual form. However, it should be noted that visemes of a single phoneme are not necessarily the same, in that a phoneme gets various shapes thanks to the influences exerted by its former and latter phonemes, altogether termed coarticulation, see Figure 1.

Viseme classification should be done based on the visual information about lip, coarticulation effects, and applications they had as a target. According to [25], if classification is provided based on acoustic data, the results could be quite different, as /m/ and /n/ which are acoustically similar, but unlike in visual appearance. Various lip shapes taken in a certain phoneme in different languages, and even in divergent accents is the reason for separately classifying visemes in them. In English, as an instance, lip height in / ãː/ is the maximum among other sounds, whereas in Persian, /æ/ sound takes this ranking.

In this study, considering coarticulation effect, those Persian visemes (CV and CVC combinations, where C stands for Consonant and V stands for Vowel) which look similar are clustered together. Without categorizing entire 29 Persian visemes, it is not possible to further reduce the processing time for the applications which would be the benefit of utilizing visemes. To this end, an agglomerative unweighted pair group method is utilized, which gives feature vector as a result of applying Interlaced Derivative Patterns (IDP) [26], concerning the effects of coarticulation and the phoneme position in syllable.

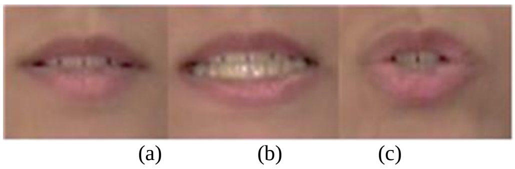


Figure 1. Different viseme for pronouncing /s/ in (a) /tas/, (b) /tes/, and (c) /tos/

IDP is a fully directional derivative pattern that takes the advantage of more detailed high-order derivative descriptions and keeps the spatial relationships in local regions. It has the capability of extracting the most promising and discriminative features which could represent data in a lower dimension phoneme space. Then a furthest neighbor of the weight value of each sample's feature vector is set as the criterion for comparing viseme dissimilarity. Afterwards, obtained clusters have been considered as the classes to which phonemes should be classified and a Support Vector Machine (SVM) is used to achieve this aim. One way to use SVM in multi-class problems is to decompose the problem into several binary ones. We decomposed our multiclass problem into k (k-1)/2 binary problems (where k is the number of classes) according to one-versus-one scheme. Therefore, each problem is addressed by means of a binary SVM which is trained in order to separate the samples of the two corresponding classes. A new sample is then classified by combining the labels predicted by these binary classifiers. Finally, we use a voting strategy for aggregating these results.

It should be noted that utilizing an appropriate data in Persian language is the first step towards viseme extraction and analysis. Therefore, we employed AVA, an audio-visual data corpus aimed for teaching sound Persian phoneme articulation [27]. On the other hand, we used AVA II database [28] because of considering coarticulation. In order to verify the algorithm's accuracy and check its conformity with the reality and with the actual human perception the state-of-the-art algorithms were compared with the proposed one.

The organization of this paper is as follows. The proposed approach comes in Section II. Section III evaluates the proposed approach, and discusses the outcome of the evaluation. Finally, Section IV concludes the paper.

II. Proposed Method

The focus of this study is on the CV and CVC combinations in Persian. The proposed approach is based on both linguistic issues and algorithmic processes. For linguistic issues, we considered phoneme position in each syllable and coarticulation effects. To make it happen, there is a need to have the required data for the following steps already prepared. Then, the agglomerative unweighted pair group method is utilized, based on which Persian visemes are clustered.

A. Linguistic Issues for Frame Selection

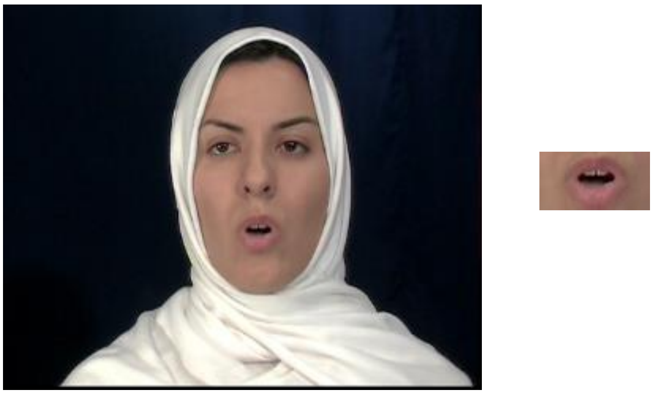
Considering linguistic issues as an algorithm prerequisite distinguishes this study from other algorithmic researches where have not fully strongly supported this viewpoint. Phoneme position in speech pattern and coarticulation effects are two important factors in visemes' appearance. Lip appearance in a speech pattern is reliant upon its place of articulation, whether it is at the beginning, in the middle or at the end; for instance, the consonant /b/ in a C_1VC_2 pattern offers different lip shapes when occurred in C_1 and C_2, respectively. The second factor is the coarticulation effect; where for example the viseme of /b/ is not the same when proceeding /u/ and /a/.

This coarticulation effect can be taken into consideration by using the middle image in bi-viseme (CV syllable) and tri-viseme (CVC syllable). Therefore, a central phoneme frame of a viseme's video sequence is manually selected for the clustering and classification tasks. In order to achieve more reliable and realistic results, a linguist actively tests out every necessary step of the selection process.

B. Lip Localization

The first step in the proposed method is to localize lip images in face images. Since the speaker has some head movements in a video sequence, and the number of the needed sequences is very large, an automatic cropping procedure is utilized to crop lip area from face.

To find the exact lip position, two criteria are considered as follows: (1) With respect to the color divergence of the nostrils to the skin color, the bottom point of the nose is detected. (2) The right and left lip corners are extracted, with respect to that the color of the lip is different from that of the skin. This work came in detail in our work on [1]. Figure 2 shows a sample of localizing the region of interest.



(a) (b)

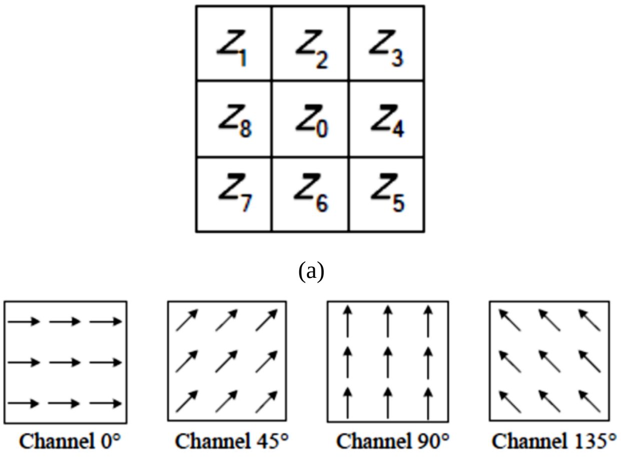
Figure 2. Extracting Region of Interest (ROI); (a) Face image (b) extracted lip

C. Feature Extraction Using Interlaced Derivative Patterns

IDP is a fully directional derivative pattern that takes the advantage of more detailed high-order derivative descriptions and keeps the spatial relationships in local regions.

In this technique, an IDP image is produced from the original image. The IDP image is a four-channel derivative image, representing four directional n^th^-order derivative channels in 0°, 45°, 90°, and 135°, respectively. The order of derivatives is derived from the order of the IDP operator; i.e., for an n^th^-order IDP operator, the IDP image with four (n-1)^th^-order derivative channels is produced. These derivative channels present more detailed description of the image in all possible directions (see Figure 3). A 3×3 neighborhood is selected around each point in the original image and the pixel is located in the IDP image. For each neighbor, the direction between the center and the neighbor is computed and the IDP image channel with the same direction is selected.

The neighbor is thresholded with the center pixel value in the selected IDP channel and the result is encoded as a binary number. This thresholding actually encodes the binary result of the first-order derivative among local neighbors and produce an extra order for the IDP operator.



(a)

(b)

Figure 3. (a) A 3×3 neighborhood around a pixel. (b) Four directional derivative channels in the IDP image.

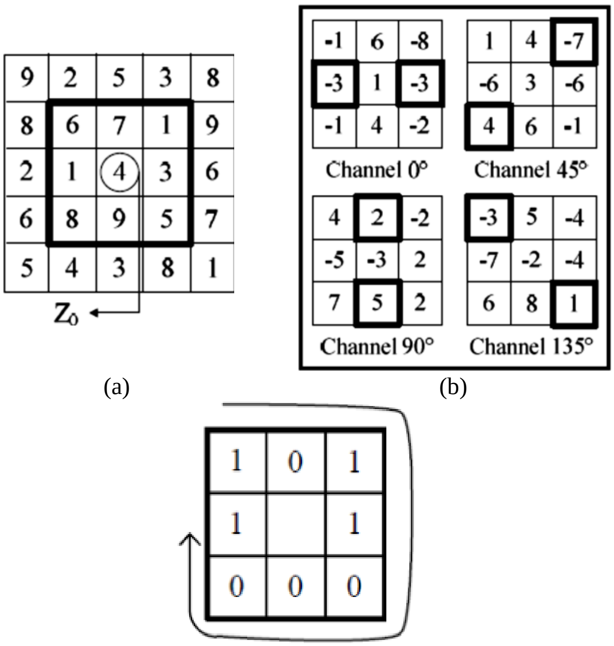
The n^th^-order IDP operator is presented in (1).

$$\begin{aligned} IDP(Z_0) = \{ & f\left(I_{135}^{n-1}(Z_0), I_{135}^{n-1}(Z_1)\right), \\ & f\left(I_{90}^{n-1}(Z_0), I_{90}^{n-1}(Z_2)\right), \\ & f\left(I_{45}^{n-1}(Z_0), I_{45}^{n-1}(Z_3)\right), \\ & f\left(I_{0}^{n-1}(Z_0), I_{0}^{n-1}(Z_4)\right), \\ & f\left(I_{135}^{n-1}(Z_0), I_{135}^{n-1}(Z_5)\right), \\ & f\left(I_{90}^{n-1}(Z_0), I_{90}^{n-1}(Z_6)\right), \\ & f\left(I_{45}^{n-1}(Z_0), I_{45}^{n-1}(Z_7)\right), \\ & f\left(I_{0}^{n-1}(Z_0), I_{0}^{n-1}(Z_8)\right) \} \end{aligned} \tag{1}$$

where the function f is defined as:

$$f(x,y) = f(x) = \begin{cases} 1, & if\,(x-y) \geq 0 \\ 0, & if\,(x-y) < 0 \end{cases} \tag{2}$$

Therefore, in each direction, only the derivatives for the center point and its neighbor point in that particular direction will be calculated.



(a) (b)

(c)

Figure 4. (a) 3×3 neighborhood in original image. (b) 4-channel IDP representation. (c) IDP code for 1 point (10110001).

This will dramatically decrease the length of the pixel representing code produced by the proposed operator compared to the Local Derivative Pattern (LDP) operator [29]. LDP keeps the extra information in a local neighborhood, while IDP encodes the relationships in the particular directions. In this way, IDP keeps only the most important information and makes the process much faster. It produces an 8-bit representation of each pixel, which makes the operator four times faster than LDP with a 32-bit representation of pixels.

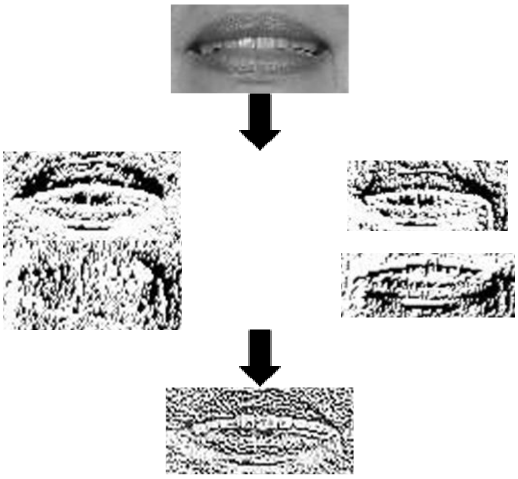


Figure 5. Visualized results of IDP code generation process for a lips image

Also compared to Local Binary Pattern (LBP), IDP contains more detailed descriptions by calculating the high-order derivative directional variations, while LBP provides first-order derivative information and is incapable of describing more detailed information. Figure 4 illustrates the 2nd-order IDP operator and Figure 5 shows the visualized results of the IDP operator on a sample lips image. To extract the discriminative IDP features of the image, the image is divided into rectangular sub-regions represented by R$_i$,…, R$_L$, and the spatial histograms are used to model the distribution of Interlaced Derivative Patterns.

Taking the spatial histograms of the sub-regions and concatenating them into an enhanced feature vector as the image descriptor is more robust against pose and illumination variations than the holistic methods [2].

D. Agglomerative Unweighted Pair Group Clustering

In order to cluster visemes in an appropriate manner, an agglomerative unweighted pair group method is utilized in which Euclidean distance is calculated for each sample x_i and sample x_j as $\|x_i - x_j\|_2$, where $i, j = 1, 2, \ldots, R$, to form the pairwise distance matrix. Initially, each sample x_i would be assigned to its own cluster C$_i$. According to algorithm which tabulated in Table 1.

Table 1. Clustering algorithm

Unweighted Pair Group Method with Average Means

```
Input:
A matrix $O$ of observed pairwise distances on $k$ taxa.

Initialization:
Assign each taxon $i$ to its own cluster $C_i$.
Let $T = t_1, \ldots, t_k$ be the set of sub-trees with one leaf.
For all $1 \le i, j \le k$, let $D[i, j] = O[i, j]$

Iteration:
while ($|T| > 1$)
{
Find two taxa $i$ and $j$ such that $D[i, j]$ is maximal.
Create a new sub-tree $t_l$ with root $l$ such that
$l$ is the parent of $i$ and $j$
$height(l) = D[i, j]/2$
Define a new cluster $C_l = C_i \cup C_j$.
For all $m \neq i, j$
        $D[l, m] = \frac{|C_i|.D[i, m] + |C_j|.D[j, m]}{|C_i| + |C_j|}$
Remove $t_i$ and $t_j$ from T and add $t_l$
}
```

E. Classification

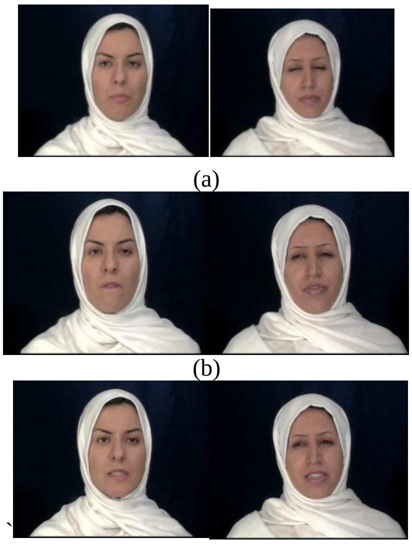
Support vector machine is one of the most powerful supervised learning methods developed by Cortes et al. [3] which separates the data space with hyperplanes or decision boundaries and is mainly designed for binary problems. One way to use SVM in multi-class problems is to decompose the problem into several binary ones. We decomposed our multiclass problem into $k(k-1)/2$ binary problems (where k is the number of classes) according to one-versus-one scheme. Therefore, each problem is addressed by means of a binary SVM which is trained in order to separate the samples of the two corresponding classes. A new sample is then classified by combining the labels predicted by these binary classifiers. Different methods are proposed for aggregating these results. A recent survey on these methods can be found in [4]. We use voting strategy for this purpose in which each binary classifier votes for a class, and the test sample is assigned to the class with the highest vote number.

III. Experimental Results

In order to demonstrate the performance of the proposed method, two sets of experiments are conducted. In the first set, IDT was compared with the LDP [5], LBP [2], Wavelet decomposition [6], and Kernel PCA with Gaussian and Polynomial kernels [7]. In the second set, results of proposed algorithm were compared with a subjective test. These methods were examined on Persian Audio/Visual data corpus [8] in order to achieve the maximum accuracy rate in the clustering and classification.

A. Data Set

Collecting a data corpus in the target language is the first step towards viseme extraction and analysis. We collected AVA, an audio-visual corpus [8] employed in this study. AVA data corpus comprises all Persian syllables, and meets the requirements of our target application. Moreover, it covers the coarticulation effect and phoneme position in syllables and sound pronunciation. The number of images processed was 2760, which came from 23×60×2, where 23 is the number of consonants in Persian, 60 is the number of image per consonant, and 2 is the number of speakers. Figure 6 shows six samples of AVA database.


(a)

(b)

(c)

Figure 6. Six samples of AVA database which show (a) /b/, (b) /v/, and (c) /s/

Table 2 tabulates Persian consonant with phoneme form, letter form in Persian, and an example.

Table 2. Persian consonant with phoneme form, letter form in Persian, and an example which consists of phonetic of the example, the example in Persian script, and its translation into English. Ph as Phoneme, Le as Letter, and Ex as Example

Ph	**Le**	**Ex**	**Ph**	**Le**	**Ex**
/p/	پ	/pedæɾ/ پدر (father)	/z/	ز، ذ، ض، ظ	/ɒːzɒːd/ آزاد (free)
/b/	ب	/bæɾɒːdæɾ/ برادر (brother)	/ʃ/	ش	/ʃɒːh/ شاه (king)
/t/	ت، ط	/tɒː/ تا (till)	/ʒ/	ژ	/ʒɒːle/ ژاله (dew)
/d/	د	/duːst/ دوست (friend)	/x/	خ	/xɒːne/ خانه (house)
/k/	ک	/keʃvæɾ/ کشور (country)	/ɢ/	غ، ق	/ɢælæm/ قلم (pen)
/g/	گ	/goruːh/ گروه (group)	/h/	ه، ح	/hæft/ هفت (seven)
/ʔ/	ء، ع	/mæʔnɒː/ معنا (meaning)	/m/	م	/mɒːdær/ مادر (mother)
/tʃ/	چ	/tʃuːb/ چوب (stick, wood)	/n/	ن	/ˈnɒːn/ نان (bread)
/dʒ/	ج	/dʒævɒːn/ جوان (young)	/l/	ل	/læb/ لب (lip)
/f/	ف	/feʃɒːɾ/ فشار (pressure)	/ɾ/	ر	/iːɾɒːn/ ایران (Iran)
/v/	و	/viːʒe/ ویژه (special)	/j/	ی	/jɒː/ یا (or)
/s/	س، ص، ث	/sɒːje/ سایه (shadow)			

B. Evaluation of Feature Extraction Method for Clustering

In order to evaluate the efficiency of IDP, the extracted features are compared the LDP, LBP, Wavelet decomposition, and Kernel PCA with Gaussian and Polynomial kernels.

LBP is defined as a grayscale invariant texture measure and is a useful tool to model texture images. LBP has shown excellent performance in many comparative studies, in terms of both speed and discrimination performance. The original LBP operator labels the pixels of an image by thresholding the 3×3 neighborhood of each pixel with the value of the central pixel and concatenating the results binomially to form a number

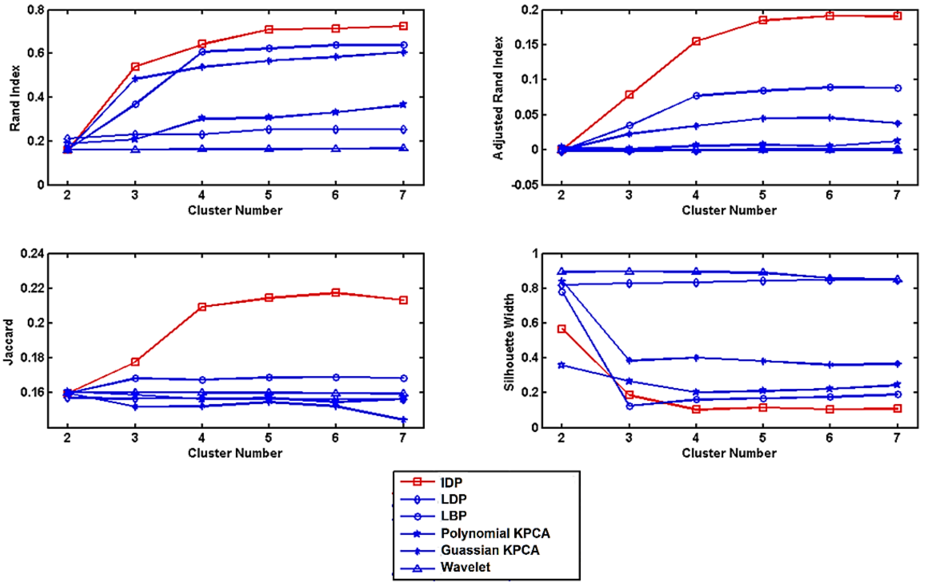


Figure 7. Comparing IDP with different feature extractor methods. Four clustering measures show the discriminative property of the IDP

. An LBP can also be considered as the concatenation of the binary gradient directions, and is called a micro pattern. The histograms of these micro patterns contain information about the distribution of the edges, spots, and other local features in an image. Different from statistic learning methods tuning a large number of parameters, the LBP method is very efficient because of its easy-to-compute feature extraction operation and simple matching strategy.

An LDP operator uses the (n-1)th-order derivative direction variations based on a binary coding function. In this scheme, LBP is conceptually regarded as the nondirectional first-order local pattern operator, because LBP encodes all-direction first-order derivative binary result, while LDP encodes the higher-order derivative information which contains more detailed discriminative features that the first-order local pattern (LBP) cannot obtain from an image.

Wavelet decomposition is worked by means of a low pass and band pass filter. The low pass filter constructs the approximate image and the band pass filter constructs detailed images.

In the conducted experiments, level two of decomposition with Haar filter is used which results in constructing a feature vector with 17 entries. The first and second entries of each feature vector relates to the mean and standard deviation of approximate image while the other entries are the standard deviation of detailed images.

Kernel PCA is a nonlinear mapping which reformulates the traditional linear PCA in a high-dimensional space using a kernel function. Kernel PCA computes the principal eigenvectors of the kernel matrix, rather than those of the covariance matrix. The reformulation of PCA in kernel space is straightforward, since a kernel matrix is constructed using the kernel function, where Gaussian and Polynomial ones are used in this work. The application of PCA in the kernel space provides Kernel PCA with the property of constructing.

Results of applying the proposed clustering approach to the extracted features show that there is a significant difference between IDP and other features based on the Rand, Adjusted Rand, Jaccard, and Silhouetted measures. This issue is evident in Figure 7.

C. Clustering Visemes Using Hierarchical Clustering Algorithm

Clustering viseme in Persian language was down through processing the records taken from two female speakers. The first speaker is aware of sound speech rules based, where the other is an ordinary prototype and is used for testing section. Results were obtained from applying the algorithm on 2760 image visemes for both speakers.

Table 3. Results of clustering Persian phonemes.

	Speaker 1	*Speaker 2*
Cluster1	/p/, /b/, /m/ /پ/, /ب/, /م/	/p/, /b/, /m/ /پ/, /ب/, /م/
Cluster2	/f/, /v/ /ف/, /و/	/f/, /v/ /ف/, /و/
Cluster3	/d/, /t/ /د/, /{ت، ط}/	/d/, /t/ /د/, /{ت، ط}/
Cluster4	/k/, /g/, /j/ /ک/, /گ/, /ی/	/k/, /g/, /j/ /ک/, /گ/, /ی/
Cluster5	/ɢ/, /n/, /h/, /x/, /ʔ/ /{غ، ق}/, /ن/, /{ح، ه}/, /خ/, /{ء، ع}/	/ɢ/, /h/, /x/, /ʔ/ /{غ، ق}/, /{ح، ه}/, /خ/, /ء، } {ع/
Cluster6	/ɾ/, /l/ /ر/, /ل/	/ɾ/, /l/, /n/ /ر/, /ل/, /ن/
Cluster7	/tʃ/, /dʒ/, /s/, /z/, /ʃ/, /ʒ/ /چ/, /ج/, /{ث، س، ص}/, /{ذ، ز، ض، ظ}/, /ش/, /ژ/	/tʃ/, /dʒ/, /s/, /z/, /ʃ/, /ʒ/ /چ/, /ج/, /{ث، س، ص}/, /{ذ، ز، ض، ظ}/, /ش/, /ژ/

.

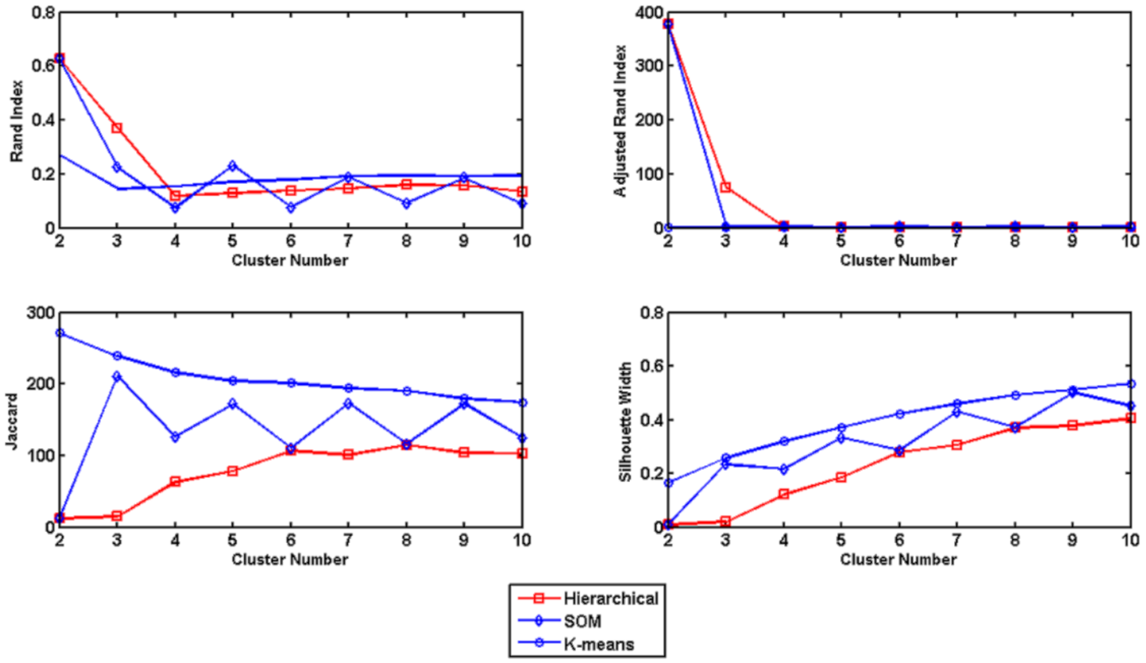


Figure 8. The Hierarchical method in comparison with K-means and SOM.

Table 3 presented the clustering results on Persian viseme. As depicted in this table, seven clusters are yielded for each speaker. In the resulting viseme groups, all visemes, except the /n/ viseme prove the same, which indicates the algorithm's fair accuracy.

The results offer that the first three groups are completely identical. The remarkable similarity between visemes stationed in every group of both testing systems is noticeable. Figure 8 shows that the proposed clustering method outperforms K-means and SOM clustering approaches regarding Rand index, Adjusted Rand index, Jaccard, and Silhouette width measures

D. Subjective Test

The subjective test can further evaluate the algorithm. In the conducted test, 30 university students who study different computer fields were randomly selected. They all had good sight and hearing abilities. Moreover, the selected viewers had not taken any previous lip reading lessons. Prior to the test implementation, the viewers are told how to answer the test, and what consonants, vowels and their equivalent symbols are. There were three issues regarding this test which were replaying the video, the speed of the played video, and speaker selection.

Since the goal is to classify consonants the movie is replayed for the second time in case the combinations were harder to understand and the majority asked for repetition. The speed of the played video can be faster/slower than the natural recorded pace. In higher speeds, the recognition rate would become more difficult or impossible occasionally, whereas at lower speeds, due to the superfluous pauses made in the middle, recognizing becomes troublesome. According to [9], the best speed for grouping ranges between $^1/_2$ and $^1/_4$ of the natural recording speed. If the speaker is an ordinary person or is unaware of standard speech techniques, incorrect articulating and mispronouncing some phonemes can be seen which in turn results in producing improper visemes. In this test, a part of the database is selected for viseme grouping, where the speech therapist utters CVC combinations. C includes all the 23 consonants in Persian language, and /e/ is used for V.

At the start of the test, the answer sheets are given among the viewers, each containing a 23×23 table, where the rows are numbered from 1 to 23, with randomly selected columns of phoneme combinations with vowel /e/. Then, the soundless test film was presented to the viewers. The viewers distinguished what was said on the video, and responded in answer sheets. The responses are collected in a confusion matrix, representing the similarity between the video items.

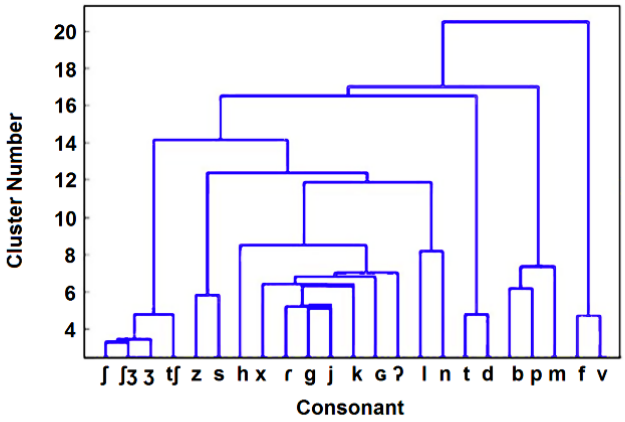


Figure 9. Dendrogram of subjective clustering results

The identified confusion rates are then categorized into viseme which is depicted in Figure 9. Finally, the subjective test result is entailed in Table 4.

As stated, the subjective test yielded 7 viseme groups. By comparing the algorithm's result (see Table 3) and the human test result (see Table 4) the algorithm's high accuracy can be noticed.

Table 4. Results of subjective clustering for Persian phonemes.

	Speaker
Cluster 1	/b/, /p/, /m/
Cluster 2	/f/, /v/

Cluster 3	/d/, /t/
Cluster 4	/s/, /z/
Cluster 5	/tʃ/, /dʒ/, /ʃ/, /ʒ/
Cluster 6	/l/, /n/
Cluster 7	/ɢ/, /h/, /x/, /ʔ/, /ɾ/, /j/, /g/, /k/

The results offer that the first three groups are completely identical. The remarkable similarity between visemes situated in every group of both testing systems is noticeable. The following visemes {/m/, /b/, /p/}, {/s/, /z/}, {/ʒ/, /ʃʒ/, /tʃ/, /ʃ/}, {/f/, /v/}, {/t/, /d/}, {/ʔ/, /ɢ/, /h/, /n/, /x/}, {/k/, /g/, /j/}, and {/ɾ/, /l/} come in the same groups. This fact demonstrates that the proposed algorithm provides a satisfactory method, the results of which are remarkably proximate to that of the actual human test. The resulted fourth and fifth groups from the subjective test, in a mixed form, are not shared with the algorithm results, in that teeth are put together in /j/, /ʒ/, /ʃʒ/, /tʃ/, and /ʃ/ visemes. Also, neglecting the role of the tongue makes some divergence. It is evident that computerized viseme grouping, because of its superior accuracy, is more reliable. This has been analyzed and confirmed by the team’s linguist/speech therapist.

E. *Classification Results*

In our experiments, we applied LDP and evaluated its feasibility to the viseme classification for the first time. We also compared the IDP approach as an improvement to LDP technique in terms of performance and efficiency with LBP and LDP in a viseme recognition task. All three techniques were used to produce appropriate features for viseme classification. In each case, the operator was applied on all images. The images were divided into 10×10 sub-regions and the image feature vector was produced by concatenating all sub-region histograms.

The recognition rate was estimated with five-fold cross validation where the SVMs’ kernels are set to linear. A pair of mean viseme was produced for each set of training subjects, and histogram intersection in Eq.6 was applied to measure the similarity between the test subject and the visemes.

$$S_{HI}(H, S) = \sum_{i=1}^{B} (H_i, S_i) \qquad (6)$$

where S_{HI} (H, S) is the histogram intersection statistic with $H = (H_1, ..., H_8)^T$ and $S = (S_1, ..., S_8)^T$.

Different orders for LDP and IDP operators were tested. We found that higher order of LDP is required for viseme classification. The 4th-order LDP performed the best in viseme recognition. In the proposed approach, the 2nd-order IDP had the highest performance and outperformed LBP and LDP.

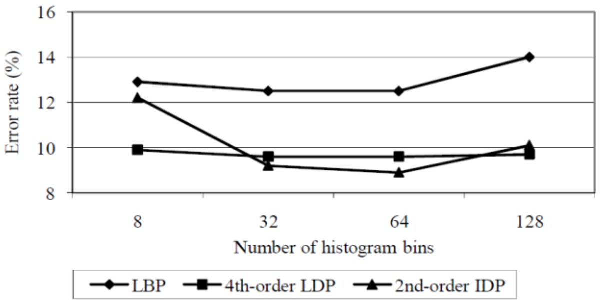


Figure 10. Classification error rates

Also, different numbers of histogram bins in each sub-region were experimented and the recognition error rate curves of all the operators remained relatively flat (see Figure 10).

Table 5 shows the error rates of the three operators and demonstrates that IDP technique outperformed LBP and LDP techniques with the highest recognition rate of 91.2%. The results show that the more detailed information extracted by IDP and LDP is more effective for viseme classification than the first-order derivative information of LBP.

As shown in Figure 10, as the number of coefficients increases, error rate gradually decreases and valleys at 8.8% with 64 IDP features. Afterwards, the rate changes slightly and stays above 10%. We can see more IDP features do not necessarily mean better recognition performance. The reason is that as the number of features increases, it added more and more to the feature space which are related to the unstable and variable details.

Table 5. Classification Error Rates of Different Techniques with 64 Histogram Bins

Method	Error rate
LBP	12.5%
LDP (4th-order)	9.6%
IDP (2nd-order)	**8.8%**

IV. Conclusion

As stated in this paper, Persian language’s visemes are clustered and classified by the accurate proposed algorithm, having speech therapy applications and photo realistic talking head animation in target. Moreover, coarticulation and phoneme position in syllables are considered. Two female respondents are captured; the first one who is aware of sound speech rules is used in viseme clustering. Based on the target application, covering coarticulation and phoneme position in syllables, a large amount of data was needed. We rationally reduced the dimensions of images by applying IDP to each of the visemes. Then the weight criterion out of the reconstruction of each viseme with the other is used for quantifying visemes’ dissimilarity through utilizing unweighted pair group method with arithmetic mean. Two evaluation procedures were considered for verifying our work. The algorithm was entirely applied to Persian speakers so as to check the robustness of the feature extraction method as well as clustering and classifying approaches compare with the state of the art methods. Comparing the results of the proposed algorithm with an expert speech therapist indicated the accuracy of our method.

References

1. R. Allahverdi, et al., “EigenCoin: sassanid coins classification based on Bhattacharyya distance,” Proc. International Conference on Information Technology, AWERProcedia Information Technology & Computer Science, 2012, pp. 1151-1160.
2. A. Bastanfard, et al., “Iranian face database with age, pose and expression,” Proc. Machine Vision, 2007. ICMV 2007. International Conference on, IEEE, 2007, pp. 50-55.
3. M.M. Dehshibi, et al., “Kernel-Based Persian Viseme Clustering,” Proc. Hybrid Intelligent Systems (HIS), 2013 13th International Conference on, 2013, pp. 130-134.
4. M.M. Dehshibi and S.M. Alavi, “Generic Visual Recognition on Non-Uniform Distributions Based on AdaBoost Codebooks,” Proc. International Conference on Image Processing, Computer Vision, and Pattern Recognition, 2011, pp. 1046-1051.
5. M.M. Dehshibi and R. Allahverdi, “Persian Vehicle License Plate Recognition Using Multiclass Adaboost,” International Journal of Computer and Electrical Engineering, vol. 4, no. 2, 2012, pp. 355-358.

6. M.M. Dehshibi and A. Bastanfard, “Unsupervised Feature Based Facial Family Similarity Recognition,” Proc. International Converence on Image and Video Processing and Computer Vision (IVPCV-10), ISRST, 2010, pp. 132-138.
7. M.M. Dehshibi and A. Bastanfard, “Portability: A New Challenge on Designing Family Image Database,” Proc. IPCV, 2010, pp. 270-276.
8. M.M. Dehshibi and A. Bastanfard, “A new algorithm for age recognition from facial images,” Signal Processing, vol. 90, no. 8, 2010, pp. 2431-2444.
9. M.M. Dehshibi, et al., “LPT: Eye Features Localizer in an N-Dimensional Image Space,” Proc. IPCV, 2010, pp. 347-352.
10. M.M. Dehshibi, et al., “Linear principal transformation: toward locating features in N-dimensional image space,” Multimedia Tools and Applications, 2013, pp. 1-25.
11. M.M. Dehshibi, et al., “Linear principal transformation: toward locating features in N-dimensional image space,” Multimedia Tools and Applications, vol. 72, no. 3, 2014, pp. 2249-2273.
12. M.M. Dehshibi, et al., “Facial family similarity recognition using Local Gabor Binary Pattern Histogram Sequence,” Proc. Hybrid Intelligent Systems (HIS), 2012 12th International Conference on, IEEE, 2012, pp. 219-224.
13. M.M. Dehshibi, et al., “Kernel-Based Object Tracking Using Particle Filter with Incremental Bhattacharyya Similarity,” Proc. Hybrid Intelligent Systems (HIS), 2013 13th International Conference on, IEEE, 2013, pp. 50-54.
14. R. Safabakhsh and F. Mirzazadeh, “AUT-Talk: a farsi talking head,” Proc. Information and Communication Technologies, 2006. ICTTA'06. 2nd, IEEE, 2006, pp. 2994-2998.
15. K. Resmi, et al., “Graphical speech training system for hearing impaired,” Proc. Image Information Processing (ICIIP), 2011 International Conference on, IEEE, 2011, pp. 1-6.
16. R.C. Luo, et al., “Human robot interactions using speech synthesis and recognition with lip synchronization,” Proc. IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society, IEEE, 2011, pp. 171-176.
17. K.-W. Yuen, et al., “Enunciate: An Internet-accessible computer-aided pronunciation training system and related user evaluations,” Proc. Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on, IEEE, 2011, pp. 85-90.
18. S. Werda, et al., “Lip localization and viseme classification for visual speech recognition,” arXiv preprint arXiv:1301.4558, 2013.
19. M. Aghaahmadi, et al., “Clustering Persian viseme using phoneme subspace for developing visual speech application,” Multimedia tools and applications, vol. 65, no. 3, 2013, pp. 521-541.
20. T. Ezzat and T. Poggio, “Visual speech synthesis by morphing visemes,” International Journal of Computer Vision, vol. 38, no. 1, 2000, pp. 45-57.
21. Z. Krňoul, et al., “Viseme analysis for speech-driven facial animation for Czech audio-visual speech synthesis,” SPECOM 2005 proceedings, 2005.
22. J. Melenchón, et al., “Objective viseme extraction and audiovisual uncertainty: estimation limits between auditory and visual modes,” Proc. AVSP, 2007, pp. 13.
23. H. Kjellström, et al., “Audio-visual phoneme classification for pronunciation training applications,” Proc. INTERSPEECH, 2007, pp. 702-705.
24. C.G. Fisher, “Confusions among visually perceived consonants,” Journal of Speech, Language, and Hearing Research, vol. 11, no. 4, 1968, pp. 796-804.
25. H. Karabalkan and H. Erdoğan, “Audio-visual speech recognition in vehicular noise using a multi-classifier approach,” 2007.
26. A. Shobeirinejad and Y. Gao, “Gender classification using interlaced derivative patterns,” Proc. Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE, 2010, pp. 1509-1512.
27. A. Bastanfard, et al., “A comprehensive audio-visual corpus for teaching sound persian phoneme articulation,” Proc. Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on, IEEE, 2009, pp. 169-174.
28. A. Bastanfard, et al., “The Persian linguistic based audio-visual data corpus, AVA II, considering coarticulation,” Advances in Multimedia Modeling, Springer, 2010, pp. 284-294.
29. B. Zhang, et al., “Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor,” Image Processing, IEEE Transactions on, vol. 19, no. 2, 2010, pp. 533-544.
30. T. Ahonen, et al., “Face description with local binary patterns: Application to face recognition,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, no. 12, 2006, pp. 2037-2041.
31. C. Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, 1995, pp. 273-297.
32. M. Galar, et al., “An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes,” Pattern Recognition, vol. 44, no. 8, 2011, pp. 1761-1776.
33. C. Garcia, et al., “A wavelet-based framework for face recognition,” Proc. Int. Workshop on Advances in Facial Image Anal. Recognition Technology, 5th European Conf. Computer Vision, Citeseer, 1998.
34. B. Schölkopf, et al., “Nonlinear component analysis as a kernel eigenvalue problem,” Neural computation, vol. 10, no. 5, 1998, pp. 1299-1319.
35. J.J. Williams, et al., “Frame rate and viseme analysis for multimedia applications to assist speechreading,” Journal of VLSI signal processing systems for signal, image and video technology, vol. 20, no. 1-2, 1998, pp. 7-23

Author Biographies

Mohammad Mahdi Dehshibi was born on 18th September 1984 in Thran, Iran. He received the B.Eng, and M. Eng in software engineering, from I.A.U, Karaj and Qazvin with honor in 2008 and 2010, respectively. He is a Ph.D candidate at I.A.U, Science and Research Branch. He has been refree of several International Conferances and Journals. His research interests lie in the field of pattern recognition, pattern formation, cellular automata, and computer vision. He has authored or co-authored over 30 papers. He is a member of IEEE Computer Society, an Associate member of DISWC.

Jamshid Shanbehzadeh received his B.S. and M.S. degree in electrical engineering from University of Tehran, Tehran, Iran in 1986 and Ph.D. degree in electrical and computer engineering from Wollongong University, Australia in 1996. He is currently an Associate Professor with the Department of Computer Engineering of Kharazmi University, Tehran, Iran. His research interests include computer vision, image retrieval, image processing, image coding, computer architecture, e-learning, e-university and e-content development. He has contributed to over 120 papers published in scientific journals or conference proceedings.